Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Near-Linear Sample Complexity for L_p Polynomial Regression

Raphael A. Meyer* Cameron Musco† Christopher Musco‡ David P. Woodruff§ Samson Zhou▽

Abstract

We study L_p polynomial regression. Given query access to a function $f : [-1, 1] \to \mathbb{R}$, the goal is to find a degree d polynomial $\widehat{q}$ such that, for a given parameter $\varepsilon > 0$,

$$\|\widehat{q} - f\|_p \leq (1 + \varepsilon) \cdot \min_{q:\deg(q)\leq d} \|q - f\|_p.$$

Here $\|\cdot\|_p$ is the L_p norm, $\|g\|_p = (\int_{-1}^{1} |g(t)|^p dt)^{1/p}$. We show that querying f at points randomly drawn from the Chebyshev measure on $[-1, 1]$ is a near-optimal strategy for polynomial regression in all L_p norms. In particular, to find $\hat{q}$, it suffices to sample $O(d \frac{\text{polylog } d}{\text{poly } \varepsilon})$ points from $[-1, 1]$ with probabilities proportional to this measure. While the optimal sample complexity for polynomial regression was well understood for L_2 and L_∞, our result is the first that achieves sample complexity linear in d and error $(1 + \varepsilon)$ for other values of p without any assumptions.

Our result requires two main technical contributions. The first concerns $p \leq 2$, for which we provide explicit bounds on the L_p *Lewis weight function* of the infinite linear operator underlying polynomial regression. Using tools from the orthogonal polynomial literature, we show that this function is bounded by the Chebyshev density. Our second key contribution is to take advantage of the structure of polynomials to reduce the $p > 2$ case to the $p \leq 2$ case. By doing so, we obtain a better sample complexity than what is possible for general p-norm linear regression problems, for which $\Omega(d^{p/2})$ samples are required.

1 Introduction

We study the problem of learning a near optimal low-degree polynomial approximation to a function $f : [-1, 1] \to \mathbb{R}$ based on as few queries $f(t_1), \ldots, f(t_n)$ to the function as possible. Studied since at least the 19th century with the work of Legendre and Gauss on least squares polynomial regression, this problem remains fundamental in statistics, computational mathematics, and machine learning. Concretely, our goal is to find a degree d polynomial $\hat{q}$ that satisfies the guarantee:

$$\|\widehat{q}(t) - f(t)\|_p \leq (1 + \varepsilon) \cdot \min_{\substack{\text{degree } d \\ \text{polynomial } q}} \|q(t) - f(t)\|_p,$$

where ε is an input accuracy parameter and $\|\cdot\|_p$ is the L_p-norm, i.e., $\|g\|_p = \left(\int_{-1}^{1} |g(t)|^p dt\right)^{1/p}$.

The problem of near-optimal polynomial approximation, visualized in Figure 1 and Figure 2, finds applications ranging from learning half-spaces [KKMS08], to solving parametric PDEs [HD15], to surface reconstruction [Pra87]. The choice of norm depends on the application: for example, $p = 1$ is used in robust approximation, $p = 2$ is common in computational science settings [CM17], and $p = \infty$ is popular in applications where f is smooth and known to admit a good minimax polynomial approximation [KKP17, Tre12]. Values of p between 2 and ∞ offer a compromise between robustness and uniform accuracy, and find applications, e.g., in the design of polynomial finite impulse response filters in signal processing [BBS94, Dum07].

The above problem is an *active learning* or *experimental design* problem since we have the freedom to choose the query locations $t_1, \ldots, t_n$. Our goal is to answer two questions:

*New York University. E-mail: ram900@nyu.edu
†University of Massachusetts Amherst. E-mail: cmusco@cs.umass.edu
‡New York University. E-mail: cmusco@nyu.edu
§Carnegie Mellon University. E-mail: dwoodruf@cs.cmu.edu
▽UC Berkeley and Rice University. Work done in part while at Carnegie Mellon University. E-mail: samsonzhou@gmail.com

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

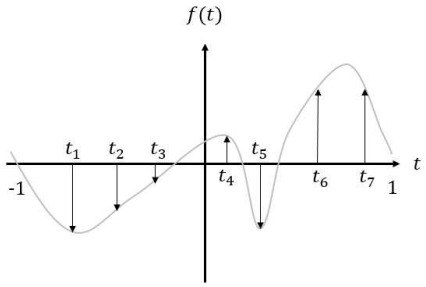

Figure 1: We choose points $t_1, \ldots, t_n$ at which to query a function f. Based on $f(t_1), \ldots, f(t_n)$, we want to find a polynomial approximating f on $[-1, 1]$.

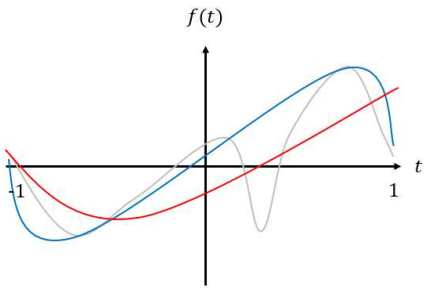

Figure 2: The blue curve is a near optimal approximating polynomial of degree 3 for $p = 2$, while the red curve is near optimal for $p = \infty$.

1. As a function of the degree d, norm p, and tolerance ε, how many queries n are required to find $\hat{q}$?
2. How should the query locations $t_1, \ldots, t_n$ be chosen from $[-1, 1]$?

When f is already a degree d polynomial, via direct interpolation, $d+1$ queries are necessary and sufficient to exactly fit f. When f is not a polynomial, we will require more than $d+1$ queries.

The above two questions have been studied extensively for $p = 2$ and $p = \infty$ [Tre12, RW12, CM17, HD15]. It is well-known that it is sub-optimal to select $t_1, \ldots, t_n$ either from an evenly spaced grid or uniformly at random: methods that try to recover $\hat{q}$ from uniform samples suffer from Runge's phenomenon [BX09, CDL13]. Improved results are obtained by selecting more queries near the *edges* of the interval $[-1, 1]$. When $p = \infty$, the typical approach is to select queries at the Chebyshev nodes [Tre12]. Classical work in approximation theory shows that, with $d+1$ samples, this approach gives an $O(\log d)$ approximation in the L_∞ norm if either polynomial interpolation or a truncated Chebyshev series is used to construct the approximation $\widehat{q}$ [Pow67, Tre12].

For $p = 2$, a recent line of work studies randomly querying according to the non-uniform Chebyshev density, which is the asymptotic density of the Chebyshev nodes:

Definition 1.1. (Chebyshev density) *For $t \in [-1, 1]$ the Chebyshev density at t is $\frac{1}{\pi\sqrt{1-t^2}}$.*

The Chebyshev density is larger for values of t near 1 and -1, and is smallest in the center of the interval, as shown in Figure 4. Prior work proves that sampling query points independently according to this density and then solving a weighted least squares problem returns a solution to the L_2 polynomial regression problem with accuracy $(1+\varepsilon)$ using $O\left(d \log d + \frac{d}{\varepsilon}\right)$ queries [RW12, CDL13, CM17]. This bound is optimal up to a $\log d$ factor: Chen and Price achieve an $O\left(\frac{d}{\varepsilon}\right)$ result using an alternative approach [CP19a], with a matching lower bound. It has also been shown that Chebyshev density sampling solves the L_∞ problem to a constant approximation factor with $O(d \log d)$ samples, improving on the $O(\log d)$ approximation guarantee for $d+1$ samples that can be obtained via classic techniques [KKP17].

In contrast to L_∞ and L_2, there have been far fewer results on near optimal polynomial regression for general p. The case of L_1 has been studied in the context of robust polynomial regression [KKP17], but results are only given under the strong assumption that f is L_∞ close to an unknown polynomial. With effort, and at the cost of a computationally expensive sampling procedure, it is possible to extend existing results on active linear regression to obtain near optimal sample complexity bounds for $p \in [1, 2]$ (see Section 1.2 for details). However, for larger values of p, all prior methods either require super-linear sample complexity ($\Omega(d^2)$ or larger), or yield a constant factor instead of a $(1+\varepsilon)$ factor approximation.

1.1 Our Contributions We give the first algorithm for active polynomial approximation that simultaneously achieves sample complexity linear in d and a $(1+\varepsilon)$ approximation factor for all L_p norms. Moreover, our procedure is simple, computationally efficient, and *universal*: we just sample points from the Chebyshev density, regardless of the value of p. That is, the same approach that works for the L_2 norm surprisingly extends to all L_p norms. Our main result is:

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

THEOREM 1.1. *For any degree d, $p \geq 1$, and accuracy parameter $\varepsilon \in (0,1)$, there is an algorithm*[1] *that queries f at $n = d\left(\frac{p\log(d)}{\varepsilon}\right)^{O(p)}$ points $t_1, \ldots, t_n$, each selected independently at random according to the Chebyshev density on $[-1,1]$, and outputs a degree d polynomial $\hat{q}(t)$ such that, with high probability,*

$$\|\widehat{q}(t) - f(t)\|_p^p \leq (1+\varepsilon) \cdot \min_{q:\deg(q)\leq d} \|q(t) - f(t)\|_p^p.$$

In addition to the simple sampling procedure, the algorithm for recovering $\widehat{q}$ is also simple: to achieve a constant factor approximation, we show that it suffices to solve an ℓ_p polynomial regression problem to find the best degree d polynomial approximating f at our queried points, reweighted appropriately[2]. To obtain a $(1+\varepsilon)$ factor approximation, we first compute a constant factor approximation $q(t)$, and then run the same regression algorithm on the residual $f(t) - q(t)$. This type of two-stage approach has been used several times in prior work on active learning for linear regression problems [DDH+08, MMWY22].

The full pseudocode is included in Algorithm 1 and Algorithm 2 below.

Algorithm 1 Chebyshev sampling for L_p polynomial approximation, Constant Factor Approximation

Input: Access to function f, parameter $p \geq 1$, degree d, number of samples n
Output: Degree d polynomial $q(t)$

1: Sample $t_1, \ldots, t_n \in [-1,1]$ i.i.d. from the pdf $\frac{1}{\pi\sqrt{1-t^2}}$
2: Observe function samples $b_i := f(t_i)$ for all $i \in [n]$
3: Build $\boldsymbol{A} \in \mathbb{R}^{n\times(d+1)}$ and diagonal $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ with $[\boldsymbol{A}]_{i,j} = t_i^{j-1}$ and $[\boldsymbol{S}]_{ii} = \left(\sqrt{1-t_i^2}\right)^{1/p}$
4: Compute $\mathbf{x} = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^{d+1}} \|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\mathbf{b}\|_p$
5: Return $q(t) = \sum_{i=1}^{d+1} x_i t^{i-1}$

Algorithm 2 Chebyshev sampling for L_p polynomial approximation, Relative Error Approximation

Input: Access to function f, parameter $p \geq 1$, degree d, number of samples n
Output: Degree d polynomial $p(t)$

1: Run Algorithm 1 on f with $\frac{n}{2}$ samples to get a polynomial $q(t)$
2: Run Algorithm 1 on $\hat{f}(t) := f(t) - q(t)$ with $\frac{n}{2}$ samples to get a polynomial $\hat{q}(t)$
3: Return $p(t) := q(t) + \hat{q}(t)$

Theorem 1.1 has a near-optimal dependence on d, since a linear dependence is required. We show that our dependence on ε is near optimal as well, proving the following lower bound:

THEOREM 1.2. *Let $p \geq 1$ be a fixed constant. Any algorithm that can output a $(1+\varepsilon)$ approximation to L_p polynomial regression with probability $\frac{2}{3}$ must use $n = \Omega(\frac{1}{\varepsilon^{p-1}})$ queries.*

It can be shown directly that no algorithm that queries f at a finite number of locations can output better than a 2-factor approximation to the best polynomial approximation in the L_∞ norm with good probability (see Section 6 or [KKP17] for details). On the other hand, a $(1+\varepsilon)$ factor approximation is achievable for $p = 2$ with just a $1/\varepsilon$ dependence in the sample complexity [CP19a]. Combined with Theorem 1.1, Theorem 1.2 helps complete the picture on the accuracy achievable for all other L_p norms.

1.2 Our Approach and Comparison to Existing Techniques Like prior work on optimal polynomial approximation in the L_2 norm [CP19a, CM17], we prove Theorem 1.1 by casting the general L_p problem as an *active linear regression problem* involving an infinitely tall design matrix (i.e., a linear operator). In the finite active linear regression problem, we are given full access to a design matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ and query access to a

[1] By an artifact of our analysis, we sample $n \sim B(n_0, \frac{m}{n_0})$ and run Algorithm 2, where $n_0 = \frac{d^6 p^{O(p)}}{\varepsilon^{O(p^2)}}$ and $m = d(\frac{p\log(d)}{\varepsilon})^{O(p)}$. This has an overall sample complexity of $d(\frac{p\log(d)}{\varepsilon})^{O(p)}$ with very high probability.

[2] We use ℓ_p to denote norms on finite dimensional spaces and L_p to denote norms on infinite dimensional spaces.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

target vector $\mathbf{b} \in \mathbb{R}^m$. The goal is to query a small number of entries from $\mathbf{b}$, and based on their values, to approximately solve $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p$.

To solve the active regression problem for $p = 2$, it is known that it suffices to sample $O(d(\log d)/\varepsilon)$ entries of $\mathbf{b}$ with probabilities proportional to the ℓ_2 *leverage scores* of the corresponding rows in $\mathbf{A}$ [Sar06]. This result generalizes to linear operators with an infinite number of rows in $\mathbf{A}$ and entries in $\mathbf{b}$ [AKM+19]. The only difference is that for linear operators, we cannot explicitly compute the ℓ_2 leverage scores (since there are an infinite number of them). To address this challenge, prior results on L_2 polynomial approximation are based on showing that, for the infinite linear operator underlying polynomial regression, the leverage scores can be tightly upper bounded by the Chebyshev measure [RW12, CM17]. Sampling by this measure thus yields an upper bound of $O(d(\log d)/\varepsilon)$ samples.

To extend these results to general L_p norms, a natural starting point is to leverage generalizations of the L_2 leverage scores to other L_p norms. There are several possible generalizations in the finite matrix case, including the ℓ_p leverage scores [DDH+08, CDW18], the ℓ_p sensitivities [CWW19, BDM+20, MMM+22], and the ℓ_p Lewis weights [CP15, CD21, PPP21, MMM+22]. Unfortunately, naïve applications of these tools to the L_p polynomial approximation problem all lead to sub-optimal guarantees. For example, it is possible to upper bound the L_p sensitivities by a scaling of the Chebyshev measure. We could then apply recent work on active regression via sensitivity sampling [MMWY22]. However, that work leads to at best a quadratic dependence on d.

Alternatively, we might hope to take advantage of recent work on active regression via sampling by ℓ_p Lewis weights – a conceptually different generalization of the ℓ_2 leverage scores than sensitivities [CD21, PPP21, MMWY22]. However, there are a few major challenges. First, we cannot explicitly compute the Lewis weights for the infinite dimensional polynomial operator, and it is much harder to obtain closed form bounds on these weights than it is for the L_2 leverage scores and L_p sensitivities. Second, for regression problems with d features, like degree-d polynomial regression, Lewis weight sampling requires $O(d^{\max(1,p/2)})$ rows [CP15, MMWY22]. So, the approach naïvely provides linear sample complexity results only for $p \in [1, 2]$.[3] For polynomial regression specifically, it is possible to use a technique from [MMM+22] to reduce from the general p case to $p \in [1, 2]$, which leads to a dp dependence, as in our Theorem 1.1. However, this reduction yields at best a constant factor approximation. The limitations of existing techniques are summarized in Figure 3.

To prove Theorem 1.1, we circumvent the above limitations for L_p Lewis weight sampling. First, for $p \leq 2$, we provide explicit bounds on the Lewis weights of the infinite linear operator underlying polynomial regression, showing that these weights are closely upper bounded by the Chebyshev measure. This almost immediately yields our results for $p \in [1, 2]$. As discussed in Section 2, doing so requires a significantly different approach than existing work on bounding leverage scores of the operator. To the best of our knowledge, our

[3] For $p \in [1, 2]$, one option would be to first carefully discretize the regression operator before computing Lewis weights, e.g., using L_p sensitivity sampling (the "first stage" in Section 2.2). While less technically involved than the $p \geq 2$ case, analyzing this approach still requires proving a bound on the L_p sensitivities of the polynomial operator. Moreover, this stage gives sub-optimal dimensionality reduction, so it would be necessary to compute the Lewis Weights of a $\tilde{O}(\frac{d^5 p^4}{\varepsilon^{2+2p}}) \times d$ matrix, using significant space and time, and resulting in a sampling procedure that is not universally good for all p.

Approach	Sample Complexity	Approximation
L_p sensitivity sampling [MMWY22]	$d^2 \left(\frac{\log d}{\varepsilon}\right)^{O(p)}$	$(1+\varepsilon)$
L_p sensitivity + Lewis weight sampling [MMWY22]	$d^{\max(1,p/2)} \left(\frac{\log d}{\varepsilon}\right)^{O(p)}$	$(1+\varepsilon)$
L_1 Lewis weight sampling [MMM+22]	$dp^2 (\log dp)^{O(1)}$	$O(1)$
Chebyshev measure sampling for all $p \geq 1$ (our results)	$d \left(\frac{\log d}{\varepsilon^p}\right)^{O(p)}$	$(1+\varepsilon)$

Figure 3: Summary of results for L_p polynomial regression. Our result is the first to obtain both an optimal linear dependence on d for all p as well as a $(1+\varepsilon)$ factor approximation.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

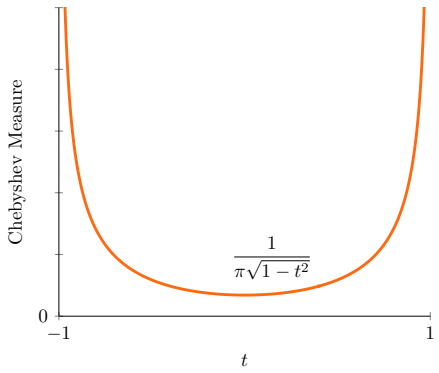

Figure 4: Plot of the Chebyshev Measure on $[-1, 1]$. Sampling from the Chebyshev measure draws fewer points from the middle of $[-1, 1]$, and more points from the ends of $[-1, 1]$

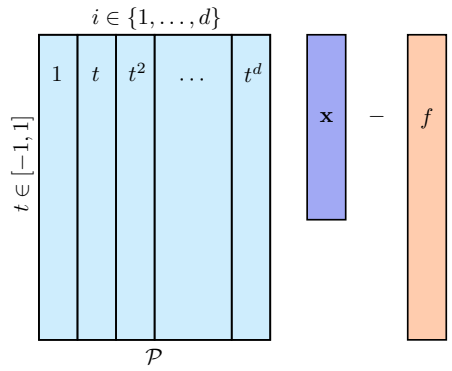

Figure 5: Visualization of the polynomial operator. $\mathcal{P}$'s column span is the set of degree d polynomials. We can approximately minimize $\|\mathcal{P}\mathbf{x} - \mathbf{f}\|_p$ by leveraging row-sampling methods for finite matrices.

bounds are the first on the Lewis weights of any natural infinite dimensional regression problem, so we hope they will be helpful in related settings where leverage scores have proven powerful. Examples include active learning for sparse Fourier functions, for bandlimited functions, and for kernel methods in machine learning [CKPS16, CP19b, AKM+19, EMM20, MM20].

Second, for $p > 2$, we need to obtain tighter bounds for Lewis weight sampling than available from black-box results that depend on $d^{p/2}$. To do so, we provide a new analysis tailored to the polynomial operator. We show that for *any* p, it actually suffices to collect $d\,\mathrm{polylog}(d)$ samples according to the Lewis weights for some other p' chosen in $[\frac{2}{3}, 2]$. Our analysis requires opening up a net analysis used in [BLM89] and [MMWY22] to analyze Lewis weight sampling for general linear operators. We leverage the fact that the $L_{p'}$ Lewis weights are close to the L_p sensitivities – both are approximated by the Chebyshev measure.

2 Technical Overview

The algorithm that achieves Theorem 1.1 is the same for all L_p norms (sample points via the Chebyshev measure and then solve two weighted ℓ_p regression problems – see Algorithm 1 and Algorithm 2). Our analysis differs for $p \in [1, 2]$ and for $p > 2$. We first describe the $p \in [1, 2]$ analysis, which is more direct.

As discussed, we solve the active polynomial approximation problem by casting it as an L_p regression problem with an infinitely tall matrix. Concretely, let $\mathcal{P} : \mathbb{R}^{d+1} \to L_2([-1, 1])$ be the *polynomial operator*, which maps a coefficient vector $\mathbf{x} \in \mathbb{R}^{d+1}$ to its corresponding degree d polynomial: $[\mathcal{P}\mathbf{x}](t) := \sum_{k=0}^{d} \mathbf{x}_k t^k$ for $t \in [-1, 1]$. Our original regression problem is equivalent to finding a vector $\hat{\mathbf{x}}$ such that

$$\|\mathcal{P}\hat{\mathbf{x}} - f\|_p^p \leq (1 + \varepsilon) \min_{\mathbf{x} \in \mathbb{R}^{d+1}} \|\mathcal{P}\mathbf{x} - f\|_p^p$$

Figure 5 visualizes this operator as matrix with infinite rows. The k^{th} column of $\mathcal{P}$ is the polynomial $t \mapsto t^k$. Each row of $\mathcal{P}$, indexed by some $t \in [-1, 1]$, is the vector $\begin{bmatrix} 1 & t & t^2 & \cdots & t^d \end{bmatrix}$.

As discussed, for $p = 2$, an effective approach to solving linear regression problems using a small number of queries of the target function is via leverage score sampling. Specifically, entries of f are sampled independently at random with probability proportional to the leverage score of the corresponding row in $\mathcal{P}$. For a finite matrix $\boldsymbol{A} \in \mathbb{R}^{m \times d}$, the leverage score of the i^{th} row of $\boldsymbol{A}$ is

$$\tau[\boldsymbol{A}](i) := \max_{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_2 > 0} \frac{([\boldsymbol{A}\mathbf{x}](i))^2}{\|\boldsymbol{A}\mathbf{x}\|_2^2}.$$

That is, $\tau[\boldsymbol{A}](i)$ is the maximum contribution that the i^{th} entry of a vector in $\boldsymbol{A}$'s range can make to its ℓ_2 norm. This definition naturally extends to linear operators [AKM+19, EMM20], and we can define

$$\tau[\mathcal{P}](t) := \max_{\mathbf{x} \in \mathbb{R}^{d+1}, \|\mathbf{x}\|_2 > 0} \frac{([\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{P}\mathbf{x}\|_2^2}.$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

For finite matrices the sum of leverage scores is always equal to the rank of $\boldsymbol{A}$, and similarly we have that $\int_{-1}^{1} \tau[\mathcal{P}](t)dt = d+1$. Recalling the particular definition of $\mathcal{P}$, we can write $\tau[\mathcal{P}](t) = \max_{\deg(q)\leq d} \frac{(q(t))^2}{\|q\|_2^2}$. It turns out that this maximum is well studied in the orthogonal polynomial literature, as it is equal to the reciprocal of the *Christoffel function* $\lambda_d(t) := \min_{\deg(q)\leq d} \frac{\|q\|_2^2}{(q(t))^2}$. While difficult to compute exactly, it can be shown that $\lambda_d(t) \geq \frac{c\sqrt{1-t^2}}{d}$ [Nev86]. This directly implies that, with appropriate scaling, the Chebyshev density upper bounds the leverage function. That is, we have $\tau[\mathcal{P}](t) \leq C\, v(t)$, where $v(t) := \frac{d+1}{\pi\sqrt{1-t^2}}$ is an appropriate scaling of the Chebyshev measure. Moreover, since $\int_{-1}^{1} v(t)dt = d+1$, we know that this upper bound is tight up to constants. Therefore, sampling from the Chebyshev density can be used to solve the L_2 polynomial approximation problem with $O(d\log d + d/\varepsilon)$ samples – only a constant factor more than would be required if sampling by the true leverage scores, which integrate to $d+1$ [CP19b].

2.1 Bounding Lewis Weights of the Polynomial Operator It has recently been shown that active regression results for finite matrices under general ℓ_p norms can be obtained by sampling by the Lewis weights, a generalization of the ℓ_2 leverage scores [CD21, MMWY22]. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times d}$, the ℓ_p Lewis weights for $\boldsymbol{A}$ are the unique numbers $w_1, \cdots, w_m$ such that

$$\tau_i(\boldsymbol{W}^{\frac{1}{2}-\frac{1}{p}}\boldsymbol{A}) = w_i \qquad \text{for all} \quad i \in [m],$$

where $\boldsymbol{W} \in \mathbb{R}^{m\times m}$ is the diagonal matrix with $\boldsymbol{W}_{ii} = w_i$. As for leverage scores, there are algorithms that compute the Lewis weights for finite matrices. But since we want to apply the weights to sample from infinite operators, it is necessary to obtain closed form bounds. It is much less clear how to do so: unlike the leverage scores, the Lewis weights are defined in a circular fashion, instead of as the solution of a natural optimization problem.

To handle this challenge, we turn to the definition of α-almost Lewis weights for matrices given in [CP15]. Specifically, we say that $w_1, \cdots, w_m$ are α-*almost* Lewis weights for $\boldsymbol{A}$ if

$$\frac{1}{\alpha} w_i \leq \tau[\boldsymbol{W}^{\frac{1}{2}-\frac{1}{p}}\boldsymbol{A}](i) \leq \alpha w_i \qquad \text{for all} \quad i \in [m] \tag{2.1}$$

where $\boldsymbol{W}$ is again the matrix with w_i on its diagonal. [CP15] prove that, for $0 < p \leq 2$, after scaling by a factor of α, the α-approximate Lewis weights are upper bounds for the true Lewis weights of a matrix.

This suggests a natural approach to bound the Lewis weights of a matrix: exhibit some weights $w_1, \cdots, w_m$ and verify that the inequality above holds. In the case of the infinite operator $\mathcal{P}$, our goal is to find a function $w(t) : [-1, 1] \to \mathbb{R}$ such that

$$\frac{1}{\alpha} w(t) \leq \tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) \leq \alpha w(t) \qquad \text{for all} \quad t \in [-1, 1] \tag{2.2}$$

where $[\mathcal{W}f](t) := w(t)f(t)$ is the linear operator equivalent to a diagonal matrix.

As a first possible candidate for the Lewis weight function, we consider the Chebyshev density $v(t)$ itself. To do so, we have to bound the leverage function $\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)$, where $[\mathcal{W}f](t) := w(t)f(t)$. We establish a surprisingly direct bound based on the fact that for each p, the weighting $\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}$ aligns with the orthogonalization measure of certain Jacobi orthogonal polynomials. Specifically, we prove:

THEOREM 2.1. *Let $J_i^{(\alpha,\beta)}(t)$ denote the degree i Jacobi Polynomial with parameters α and β. Then, letting $\alpha = \beta = \frac{1}{p} - \frac{1}{2}$, we have*

$$\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) = \frac{1}{(1-t^2)^{\frac{1}{2}-\frac{1}{p}}} \sum_{i=0}^{d} (J_i^{(\alpha,\beta)}(t))^2$$

That is, we can *exactly* characterize this Chebyshev-reweighted leverage function in terms of Jacobi polynomials. Further, because Jacobi polynomials are well studied in the orthogonal polynomial literature, we can appeal to prior work on uniformly upper bounding these polynomials to bound the above sum of squares. Overall, in Section 4 we prove:

$$\frac{1}{\alpha} v(t) \leq \tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) \leq \alpha v(t) \qquad \text{for all} \quad |t| \leq 1 - \frac{1}{d^2} \tag{2.3}$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

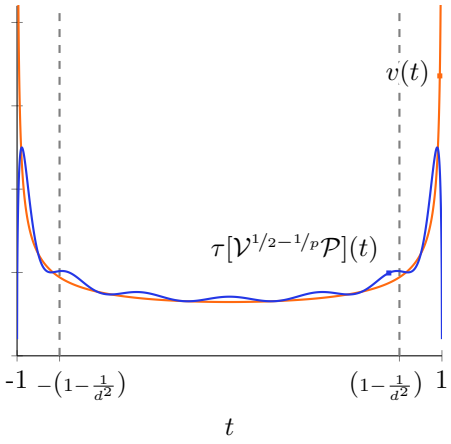

Figure 6: Plot of the scaled Chebyshev Measure (—) and corresponding reweighted leverage function $\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)$ (—) on $[-1,1]$ for $d=6$, $p=1$. For most values of t both curves are close, but for $|t|>1-\frac{1}{d^2}$ the curves diverge. This means that the Chebyshev density itself does not directly approximate the L_p Lewis weights, motivating our study of a clipped version of the measure, denoted $w(t)$.

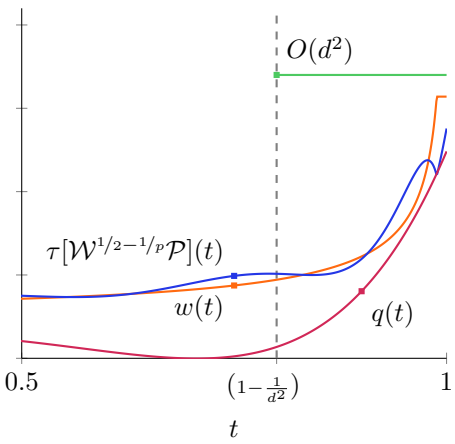

Figure 7: Plot of the clipped Chebyshev Measure (—) and corresponding reweighted leverage function (—) for $t\in[0.5,1]$ and $d=6$, $p=1$. As proven in Theorem 2.2, these functions are within a constant factor for all t, so the clipped measure approximates the L_p Lewis weights for $p\le 2$. We also visualize the "spike" polynomial $q(t)$ (—) and the upper bound (—) used in the proof of Theorem 2.2.

This is very close to what we need to show, but unfortunately the almost-Lewis weight property does not hold for large $|t|>1-\frac{1}{d^2}$. Figure 6 shows what goes wrong: the Chebyshev density $v(t)$ diverges to $+\infty$ while the weighted leverage function $\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)$ remains bounded. To resolve this issue, we adjust our proposed Lewis weight function, and instead consider $w(t):=\max\{c_1(d+1)^2,v(t)\}$, which clips the Chebyshev density so that it cannot diverge to $+\infty$. We can then show the following core theorem for small p:

THEOREM 2.2. *There are fixed constants c_1,c_2,c_3 such that, letting $w(t)=\min\left\{c_1(d+1)^2,v(t)\right\}$ be the clipped Chebyshev measure on $[-1,1]$ and letting $\mathcal{W}$ be the corresponding diagonal operator with $[\mathcal{W}f](t)=w(t)\cdot f(t)$, for any $p\in[\frac{2}{3},2]$ and $t\in[-1,1]$,*

$$\frac{c_2}{\log^3 d}w(t)\le\tau[\mathcal{W}^{1/2-1/p}\mathcal{P}](t)\le c_3 w(t)$$

Theorem 2.2 shows that the clipped Chebyshev density gives a set of $O(\log^3 d)$-almost Lewis weights for the polynomial operator. So we can upper bound the true Lewis weights by the clipped measure, and only gain a polylog(d) factor in the final sample complexity in comparison to exact Lewis weight sampling. Moreover, we can obtain the same bound via sampling by the Chebyshev measure itself, which tightly upper bounds the clipped measure after scaling (i.e., it has the same integral on $[-1,1]$ up to a constant factor). We also reiterate that when $p\in[1,2]$ we will directly appeal to this theorem for this value of p, but when $p>2$ we will appeal to this theorem for a different $p'\in[\frac{2}{3},2]$, which is why we prove Theorem 2.2 for some values of $p<1$.

We prove Theorem 2.2 by separately considering the case when $|t|\le 1-\frac{1}{d^2}$ and when $|t|>1-\frac{1}{d^2}$. The first case is easier: we show that for such values of t, the reweighted leverage function corresponding to the clipped Chebyshev measure – i.e. $\tau[\mathcal{W}^{1/2-1/p}\mathcal{P}](t)$ – very closely approximates the reweighted leverage function corresponding to the unclipped measure. We can then directly appeal to Equation 2.3. The second case is more challenging: when $|t|>1-\frac{1}{d^2}$, the density at t is different in the clipped and unclipped measure, so the reweighted leverage scores differ significantly. To deal with this hard case, we separately prove an upper and lower bound as follows:

Upper Bound: Because $w(t)$ itself is bounded, we can bound $\tau[\mathcal{W}^{1/2-1/p}\mathcal{P}](t)\le\tau[\mathcal{P}](t)$, and we use the Markov Brothers' inequality to bound $\tau[\mathcal{P}](t)\le O(d^2)$.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Lower Bound: Because $\tau[\mathcal{W}^{1/2-1/p}\mathcal{P}](t)$ is a maximization over degree d polynomials, we can prove a lower bound by exhibiting a specific "spike" polynomial $q(t)$ which has $([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}q](t))^2/\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}q\|_2^2 = \Omega(\frac{d^2}{\log^3 d})$.

The detailed proof can be found in Section 4. The final result of Theorem 2.2 is visualized in Figure 7.

2.2 Active L_p Regression via Chebyshev Sampling Now that we have now bounded the L_p Lewis weight function of the polynomial operator $\mathcal{P}$ by the Chebyshev density for $p \in [\frac{2}{3}, 2]$, in order to prove Theorem 1.1 for $p \in [1, 2]$, we can almost directly apply existing Lewis weight sampling guarantees for active ℓ_p regression [MMWY22, CD21]. However, there remains an outstanding challenge. Naïve Lewis weight sampling for ℓ_p regression on an $m \times d$ matrix incurs a $\log(m)$ dependence in the sample complexity[4]. This rules out directly applying Lewis weight sampling to our infinite operator $\mathcal{P}$, for which m is infinite (recall Figure 5).

We address this challenge with a simple observation: sampling rows of $\mathcal{P}$ by the Chebyshev measure is essentially equivalent[5] to collecting a large *uniform sample* of rows of $\mathcal{P}$ and then subsampling those rows according to the Chebyshev measure. We visualize this "two-stage" decomposition of our sampling method in Figure 8, and emphasize that we do not algorithmically generate the first uniformly sampled matrix[6]. Instead, so long as this hypothetical two-stage algorithm is correct, by the equivalence of these sampling schemes, we know that our actual algorithm is correct.

Proving correctness requires two key ingredients. Let $\boldsymbol{A} \in \mathbb{R}^{n_0 \times d+1}$ be this matrix created by uniformly sampling n_0 rows of $\mathcal{P}$. First, we show that taking $n_0 = \frac{\text{poly}(d)}{\text{poly}(\varepsilon)}$ suffices to recover a $(1+\varepsilon)$ error solution to the full regression problem on $\mathcal{P}$. Second, we prove that the Chebyshev measure evaluated at $\boldsymbol{A}$'s rows tightly upper bounds $\boldsymbol{A}$'s Lewis weight distribution. So, by prior work [MMWY22, CD21], this can be used to show that sampling by the measure suffices to obtain a $(1+\varepsilon)$ error solution to the regression problem involving $\boldsymbol{A}$. This Lewis weight sampling stage only has a dependence on $\log(n_0) = \log(\frac{d}{\varepsilon})$, avoiding the $\log(m)$ issue. Overall, combining the error guarantees of both stages ensures that our hypothetical two-stage algorithm samples rows of $\mathcal{P}$ in the same way as Algorithm 1 and with the same sample complexity as Theorem 1.1.

To prove the first point, that uniform sampling a large number of rows preserves a near-optimal solution, we turn to a different tool from the matrix sampling literature: L_p *sensitivity sampling*. The L_p sensitivities are a natural generalization of the L_2 leverage scores, defined as

$$\psi_p[\mathcal{P}](t) := \max_{\mathbf{x}\in\mathbb{R}^{d+1}} \frac{|[\mathcal{P}\mathbf{x}](t)|^p}{\|\mathcal{P}\mathbf{x}\|_p^p} = \max_{\deg(q)\le d} \frac{|q(t)|^p}{\|q\|_p^p}$$

The value of using L_p sensitivity sampling is that standard concentration bounds and an ε-net argument show that sampling $n_0 = \frac{\text{poly}(d)}{\text{poly}(\varepsilon)}$ rows proportionally to their sensitivities suffices to recover a $(1+\varepsilon)$ error solution to the full L_p regression problem. While the dependence on d is polynomially worse than that of Lewis weight sampling, it has no dependence on m. Since we want to sample rows of $\mathcal{P}$ uniformly, we will need to show a uniform bound on $\psi_p[\mathcal{P}](t)$ (i.e., an upper bound that does not depend on t). Using a classical result on the smoothness of polynomials (specifically the Markov brothers' inequality), we can indeed show $\psi_p[\mathcal{P}](t) \le d^2(p+1)$, which in turn implies that $n_0 = \frac{\text{poly}(d)}{\text{poly}(\varepsilon)}$ uniform samples suffice.

To prove the second point, we need to show that the Chebyshev measure upper bounds $\boldsymbol{A}$'s Lewis weights. To do so, we prove that the clipped Chebyshev measure, which is an almost-Lewis weight measure for $\mathcal{P}$, is also an almost-Lewis weight distribution for $\boldsymbol{A}$. Again the proof mostly follows from standard concentration results combined with an ε-net argument, although we also need to use the fact that the clipped Chebyshev measure is bounded.

We visualize the structure of our two-stage proof in Figure 8. Overall, the arguments above complete the analysis of L_p polynomial regression for $p \in [1, 2]$.

[4]Some work on Lewis weight sampling, including by Cohen and Peng [CP15], implicitly assumes $\log m = O(\log d)$. This is reasonable in the finite matrix setting, but does not apply when m is infinite.

[5]Two subtleties emerge here. First, we say "essentially equivalent" since this two-stage sampling scheme is $O(\frac{1}{\text{poly}(d)})$ close to our actual Chebyshev sampling in total variation, so these schemes are indistinguishable but not the same. Second, analyzing the two-stage procedure will require a random choice of the sample number n – see the footnote on Theorem 1.1.

[6]In principal, we *could* algorithmically generate the uniform subsampled matrix and numerically compute its ℓ_p Lewis weights, although this would incur a much higher polynomial runtime dependence on d than our simpler approach of sampling directly from the Chebyshev measure.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

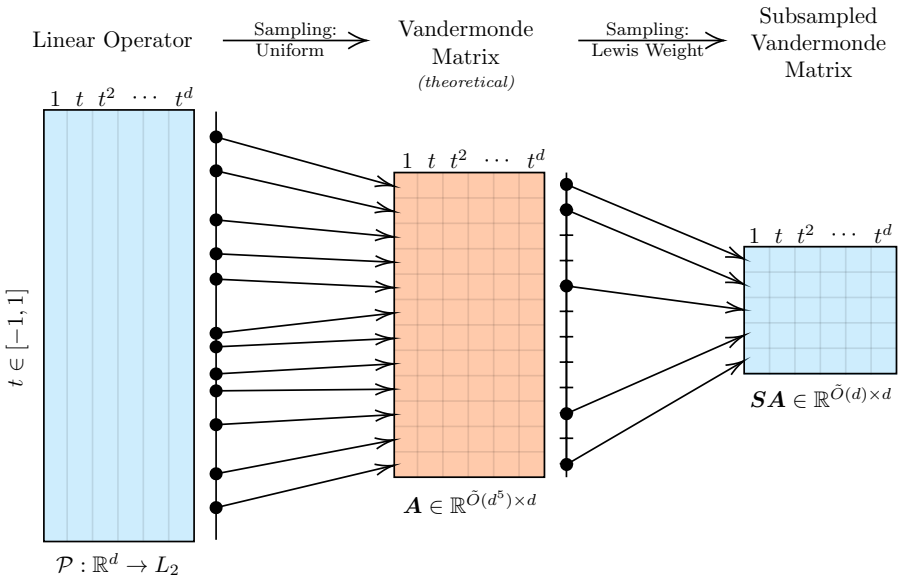

Figure 8: Sketch of the two-stage proof technique described in Section 2.2. We show that the Chebyshev measure sampling of Algorithm 1 is equivalent to a hypothetical two stage sampling procedure that first uniformly samples $O\left(\frac{\text{poly}(d)}{\text{poly}(\varepsilon)}\right)$ query points from $[-1,1]$ to form Vandermonde matrix $\boldsymbol{A}$, and then further samples the rows of $\mathbf{A}$ by the Chebyshev measure, which approximates $\mathbf{A}$'s Lewis weight distribution. Since we can uniformly bound the L_p sensitivities of the original regression problem by $\text{poly}(d)$, we can argue that both stages of sampling preserve the solution of the L_p regression problem, and thus that our final solution gives a $(1+\varepsilon)$ approximation to the optimal.

2.3 Near-Linear Sample Complexity for $p > 2$ The next challenge is to extend our results to $p > 2$. We could use a similar approach as in Section 2.1 and Section 2.2, but doing so would lead to suboptimal sample complexity. In particular, ℓ_p matrix Lewis weight sampling algorithms have a very different sample complexity for $p \leq 2$ and $p > 2$. For $p \in [0, 2]$, Lewis weight sampling requires $\tilde{O}(d)$ samples. For $p > 2$, Lewis weight sampling requires $\tilde{O}(d^{p/2})$ samples, and there are worst-case matrices that necessitate this sample complexity. So to achieve $\tilde{O}(d)$ sample complexity, we require a novel analysis of ℓ_p Lewis weight sampling for active regression that leverages the structure of the polynomial operator $\mathcal{P}$. Concretely, within the framing of Section 2.2, we keep the uniform sensitivity sampling stage but provide a new analysis for the second Lewis weight sampling stage.

We start by describing a simple approach for achieving constant factor error (but not $(1+\varepsilon)$ factor) which follows from an observation in [MMM+22]. In particular, if we only want constant factor error, it suffices to find a subsampling matrix $\boldsymbol{S}$ that satisfies an ℓ_p *subspace embedding* property. Specifically, we need that for all $\mathbf{x} \in \mathbb{R}^{d+1}$, $\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}\|_p^p \approx \|\boldsymbol{A}\mathbf{x}\|_p^p$. We argue that such a matrix can be constructed with a number of rows linear in d (for any constant p) as follows: Let f be a degree d polynomial, and let r be an integer such that $q := \frac{p}{r} \in [1, 2]$. Then, $t \mapsto (f(t))^r$ is some degree rd polynomial. So, if $\boldsymbol{A} \in \mathbb{R}^{n_0 \times d+1}$ is our Vandermonde matrix resulting from uniform sampling, we can let $\boldsymbol{B} \in \mathbb{R}^{n_0 \times rd+1}$ be another Vandermonde matrix generated by the same time samples but describing polynomials of degree rd. Then, for all $\mathbf{x} \in \mathbb{R}^{d+1}$ there exists some $\mathbf{y} \in \mathbb{R}^{rd+1}$ such that $(\boldsymbol{A}\mathbf{x})^r = \boldsymbol{B}\mathbf{y}$, where we define the exponentiation elementwise. In particular, we have $\|\boldsymbol{A}\mathbf{x}\|_p^p = \|\boldsymbol{B}\mathbf{y}\|_q^q$. Therefore, if we know that $\boldsymbol{S}$ provides an ℓ_q norm subspace embedding for $\boldsymbol{B}$, so that $\|\boldsymbol{S}\boldsymbol{B}\mathbf{y}\|_q^q \approx \|\boldsymbol{B}\mathbf{y}\|_q^q$ for all $\mathbf{y} \in \mathbb{R}^{rd+1}$, we also know that $\boldsymbol{S}$ is a subspace embedding for $\boldsymbol{A}$: $\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}\|_p^p \approx \|\boldsymbol{A}\mathbf{x}\|_p^p$ for all $\mathbf{x} \in \mathbb{R}^{d+1}$. Since $\boldsymbol{B}$ is exactly the Vandermonde matrix we would have generated from uniformly sampling in Section 2.2 with degree rd and ℓ_q norm, we know that the Chebyshev measure bounds the Lewis weights of $\boldsymbol{B}$, and that the Lewis weight subsampling matrix $\boldsymbol{S}$ is a subspace embedding for $\boldsymbol{B}$, and therefore also for $\boldsymbol{A}$.

Achieving $(1+\varepsilon)$ error regression is harder but takes a similar approach. In order to have Lewis weight sampling imply $(1+\varepsilon)$ error regression, a subspace embedding does not suffice and a more detailed argument is needed [MMWY22]. A crucial step in this analysis is showing an *affine embedding*: that $\|\boldsymbol{S}(\boldsymbol{A}\mathbf{x}-\mathbf{b})\|_p \approx \|\boldsymbol{A}\mathbf{x}-\mathbf{b}\|_p$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

for all $\boldsymbol{A}\mathbf{x}$ with small ℓ_p norm. [BLM89] and [MMWY22] provide a way to prove this affine embedding via a *compact rounding argument*, which designs a structured set of ε-nets which allow for a tight $\tilde{O}(d^{\max\{1,p/2\}})$ sample complexity to be obtained from Lewis weight sampling. To obtain a linear dependence in d for all p, we reduce from the ℓ_p case to the ℓ_q case for $q \leq 2$, as discussed above, but in a less direct way. In particular, we show that a compact rounding for the range of $\boldsymbol{B}$ can be directly transformed to construct a compact rounding of the same size for the range of $\boldsymbol{A}$.

This approach is elaborated on in Section 5.3. Critically, we will now enforce that r is also an odd integer, so that we not only get $(\boldsymbol{A}\mathbf{x})^r = \boldsymbol{B}\mathbf{y}$ but also have $\boldsymbol{A}\mathbf{x} = (\boldsymbol{B}\mathbf{y})^{1/r}$. This does not hold when r is even since negative entries of $\boldsymbol{A}\mathbf{x}$ get turned positive. For $p \geq 3$, we let r be the largest odd integer smaller than p, so that $q = \frac{p}{r} \in [1, 2]$. For $p \in (2, 3)$, this would pick $r = 1$ which would not be helpful, so we instead take $r = 3$, so that $q = \frac{p}{r} \in [\frac{2}{3}, 1]$. Once we construct this compact rounding, we find that sampling the rows of $\boldsymbol{A}$ by the ℓ_q Lewis weights of $\boldsymbol{B}$ achieves the affine embedding with sample complexity linear in d. And since Section 2.1 bounds the Lewis weights of $\boldsymbol{B}$ by the Chebyshev measure, we conclude that Algorithm 2 achieves Theorem 1.1 for all $p > 2$.

2.4 Lower Bounds and L_∞ Polynomial Approximation The linear dependence on d in Theorem 1.1 cannot be improved: when f is exactly equal to a degree d polynomial, if we do not take at least $d + 1$ samples it is not possible to recover a zero-error approximation to the function. A natural question is if the $1/\varepsilon^{O(p)}$ dependence in the theorem is also tight – i.e., is it necessary for the accuracy to depend exponentially on p?

We answer this question in the affirmative with the lower bound of Theorem 1.2, which has a short and direct proof. For any algorithm that queries f at most $n \leq O(\frac{1}{\varepsilon^{p-1}})$ times, there must exist an interval $\mathcal{I} \subset [-1, 1]$ of width ε^{p-1} such that none of the algorithm's queries lie in $\mathcal{I}$ with probability $\frac{2}{3}$. We then randomly select a function f that is either $+1$ or -1 on $\mathcal{I}$ with equal probability, and 0 elsewhere. To obtain a $1 + \varepsilon$ approximation in the L_p norm, the algorithm must distinguish between these two cases, but with probability $\frac{2}{3}$, it does not even obtain a sample from the non-zero region.

Finally, we note that our techniques can be extended to give a constant factor approximation to the L_∞ polynomial approximation problem with $O(d \operatorname{polylog}(d))$ samples. Details are discussed in Section 6, where we relate the L_∞ problem to the L_p problem with $p = O(\log d)$. Results for the L_∞ norm were already shown in [KKP17] using a different approach but the same Chebyshev measure sampling distribution.

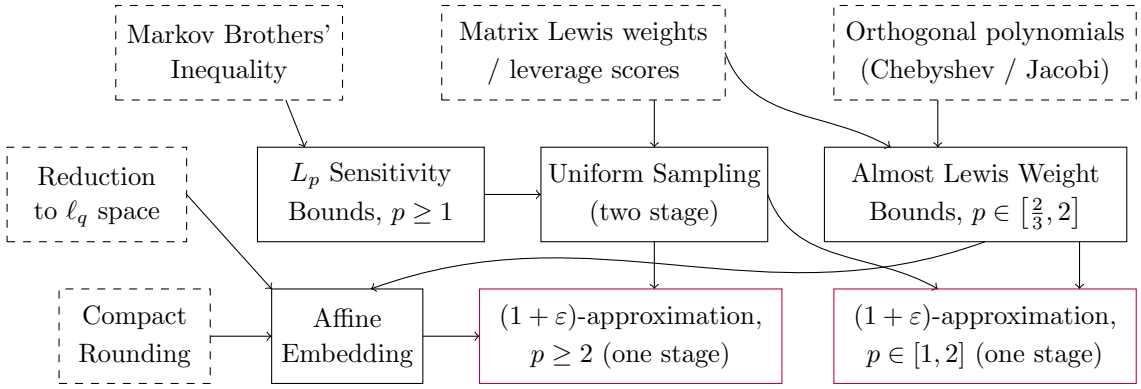

Figure 9: Flowchart of proofs: dashed rectangles represent existing results, solid rectangles represent our technical contributions.

Organization of the rest of the paper. We first consider the L_p regression problem for $p \in [1, 2]$ in Section 4. Specifically, we start by relating the Chebyshev density to the L_p Lewis weights for all $p \in [\frac{2}{3}, 2]$. We first outline the proof for $p = 1$ in Section 4.2 and defer the proof for general $p \in [\frac{2}{3}, 2]$ to Section 4.3 and Section 7. We then prove correctness of Algorithm 1 for $p \in [1, 2]$ in Section 4.4 and Section 4.5.

We handle $p > 2$ in Section 5. We first prove the correctness of constant-factor regression in Section 5.1, prove the majority of $(1 + \varepsilon)$ error analysis in Section 5.2, and prove a core technical claim for $p > 2$ in Section 5.3. We present the lower bound Theorem 1.2 in Section 5.2. Finally, we address L_∞ regression in Section 6. A summary of our high-level ideas and their dependencies is shown in Figure 9.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

3 Preliminaries

For an integer $n > 0$, we use $[n]$ to denote the set $\{1, \ldots, n\}$. We use $\text{poly}(n)$ to denote a constant degree polynomial in n and $\text{polylog}(n)$ to denote a polynomial in $\log n$.

Throughout this paper, unbold lowercase letters are scalars or functions, bold lowercase letters are vectors, bold uppercase letters are matrices, and calligraphic uppercase letters are linear operators. The norm $\|\cdot\|_p$ will interchangeably refer to the vector norm, defined by $\|\mathbf{x}\|_p^p = \sum_{i=1}^d |x_i|^p$, and the continuous norm $\|f\|_p^p = \int_{-1}^1 |f(t)|^p \, dt$. We say that a matrix $\boldsymbol{A}$ is a subspace embedding for another matrix or linear operator $\mathcal{A}$ if for all $\mathbf{x}$ we have $\frac{1}{\alpha}\|\mathcal{A}\mathbf{x}\|_p^p \leq \|\boldsymbol{A}\mathbf{x}\|_p^p \leq \alpha\|\mathcal{A}\mathbf{x}\|_p^p$ for some constant $\alpha \geq 1$. More broadly, if two scalars x and y have $\frac{1}{\alpha}x \leq y \leq \alpha x$, then we write $x \approx_\alpha y$. For instance, the subspace embedding guarantee can be written as $\|\mathcal{A}\mathbf{x}\|_p^p \approx_\alpha \|\boldsymbol{A}\mathbf{x}\|_p^p$ for all $\mathbf{x}$. We use brackets for indexing on both vectors and functions.

The i^{th} entries of the vectors $\mathbf{x}$ and $\boldsymbol{A}\mathbf{x}$ are denoted $\mathbf{x}(i)$ and $[\boldsymbol{A}\mathbf{x}](i)$. The ℓ_2 leverage score of the i^{th} row of matrix $\boldsymbol{A}$ is denoted $\tau[\boldsymbol{A}](i)$. The ℓ_p Lewis weight of the i^{th} row of matrix $\boldsymbol{A}$ is denoted $w_p[\boldsymbol{A}](i)$. The ℓ_p sensitivities of the i^{th} row of matrix $\boldsymbol{A}$ is denoted $\psi_p[\boldsymbol{A}](i)$. We similarly denote the leverage function, Lewis weight function, and sensitivity of an operator $\mathcal{A}$ at time t as $\tau[\mathcal{A}](t)$, $w_p[\mathcal{A}](t)$, and $\psi_p[\mathcal{A}](t)$.

Let $\mathcal{P}$ denote the polynomial operator of degree d:

$$\mathcal{P} : \mathbb{R}^{d+1} \to L_p([-1,1]) \qquad\qquad [\mathcal{P}\mathbf{x}](t) := \sum_{i=0}^{d} x_i t^i$$

Note that the maximum degree of a polynomial is d, but that the *rank* of $\mathcal{P}$ is $d+1$ because of the constant degree-0 polynomial.

We recall the Markov brothers' inequality that bounds the magnitude of the derivative of a polynomial of degree d whose magnitude inside the interval $[-1,1]$ is bounded by 1.

THEOREM 3.1. (MARKOV BROTHERS' INEQUALITY, E.G., THEOREM 2.1 IN [GM99]) *Suppose $q(t)$ is a polynomial of degree at most d such that $|q(t)| \leq 1$ for $t \in [-1,1]$. Then for all $t \in [-1,1]$, $|q'(t)| \leq d^2$.*

Throughout this paper, we will be analyzing Algorithm 1, and showing that this algorithm satisfies Theorem 1.1.

4 Active L_p Regression for $p \in [1,2]$

In this section, we start with the definition of leverage scores and prove that the L_1 Lewis weights for the polynomial operator are bounded by the Chebyshev measure. In particular, this section shows the relationship between Lewis weights and uniform bounds on orthogonal polynomials. We then use this Lewis weight bound to show that $\tilde{O}(d)$ samples suffice for robust L_1 regression.

4.1 Warm Up: Bounding the Leverage Scores for $p = 2$

We first start with leverage scores, which are a key building block underpinning Lewis weights. Before discussing Lewis weights, we will look at bounding the leverage scores of $\mathcal{P}$, which relates to solving L_2 regression. We first look at the properties of Leverage Scores for matrices:

DEFINITION 4.1. *For a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, the leverage score for row $i \in [n]$ is:*

$$\tau[\boldsymbol{A}](i) := \max_{\mathbf{x} \in \mathbb{R}^d, \|\boldsymbol{A}\mathbf{x}\|_2 > 0} \frac{([\boldsymbol{A}\mathbf{x}](i))^2}{\|\boldsymbol{A}\mathbf{x}\|_2^2}$$

The leverage scores of a matrix are well studied, and we will rely on two of their properties:

1. Leverage Scores are invariant to change of basis: for full-rank $\boldsymbol{U} \in \mathbb{R}^{d \times d}$, we have $\tau[\boldsymbol{A}\boldsymbol{U}](i) = \tau[\boldsymbol{A}](i)$.
2. If $\boldsymbol{A}$ has orthonormal columns, then $\tau[\boldsymbol{A}](i) = \|\mathbf{a}_i\|_2^2$ where $\mathbf{a}_i$ is the i^{th} row of $\boldsymbol{A}$.

So, if we can find a matrix $\boldsymbol{U}$ such that $\boldsymbol{A}\boldsymbol{U}$ has orthonormal columns, then we can compute $\tau_i(\boldsymbol{A}) = \|[\boldsymbol{A}\boldsymbol{U}](i)\|_2^2$. We can use this argument to bound the *Leverage Function* of the polynomial operator:

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

DEFINITION 4.2. *For an operator $\mathcal{A} : \mathbb{R}^{d+1} \to L_2([-1,1])$, the leverage function for $\mathcal{A}$ at time $t \in [-1,1]$ is*

$$\tau[\mathcal{A}](t) := \max_{\mathbf{x}\in\mathbb{R}^{d+1},\, \|\mathcal{A}\mathbf{x}\|_2>0} \frac{([\mathcal{A}\mathbf{x}](t))^2}{\|\mathcal{A}\mathbf{x}\|_2^2}$$

We can easily see that the leverage function is also rotationally invariant. As shown in Figure 5, $\mathcal{P}$ has columns that represent the first degree d monomials. That is, we think of the i^{th} column of $\mathcal{P}$ as the polynomial $p_i(t) = t^{i-1}$. Since $\int_{-1}^{1} p_i(t)p_j(t)dt \neq 0$ in general, these columns are not orthogonal.

While the first degree d monomials are not orthogonal, the Legendre polynomials are. So, we can find a change-of-basis matrix $\boldsymbol{U}$ such that the columns of $\mathcal{P}\boldsymbol{U}$ are Legendre polynomials instead. Under this basis, we have $\|\mathcal{P}\boldsymbol{U}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$, which lets us simplify the leverage function. Letting $L_i(t)$ denote the degree i Legendre polynomial, normalized so that $\int_{-1}^{1}(L_i(t))^2dt = 1$, we have

$$\tau[\mathcal{P}](t) = \max_{\mathbf{x}\in\mathbb{R}^{d+1}} \frac{([\mathcal{P}\boldsymbol{U}\mathbf{x}](t))^2}{\|\mathcal{P}\boldsymbol{U}\mathbf{x}\|_2^2} = \max_{\|\mathbf{x}\|_2=1} ([\mathcal{P}\boldsymbol{U}\mathbf{x}](t))^2 = \max_{\|\mathbf{x}\|_2=1} \left(\sum_{i=0}^{d} x_i L_i(t)\right)^2 = \sum_{i=0}^{d}(L_i(t))^2 \tag{4.4}$$

The last equality follows because $\max_{\|\mathbf{x}\|_2=1}(\mathbf{a}^\top\mathbf{x})^2 = \|\mathbf{a}\|_2^2$ for any $\mathbf{a}$. If we view $\mathcal{P}\boldsymbol{U}$ as an infinite matrix whose rows correspond to $t \in [-1,1]$ and whose columns correspond to the Legendre polynomials, then Equation 4.4 shows that $\tau[\mathcal{P}](t)$ equals the row-norm-squared of this matrix, matching the second property we mentioned for matrix leverage scores.

So, to bound the leverage function for $\mathcal{P}$, we now need to bound the sum-of-squared Legendre polynomials. Here we appeal to existing uniform bounds on orthogonal polynomials. For instance, Lorch proved in 1983 that $|L_i(t)| \leq \sqrt{\frac{2}{\pi\sqrt{1-t^2}}}$ for all $t \in [-1,1]$ [Lor83]. So we conclude the bound

$$\tau[\mathcal{P}](t) = \sum_{i=0}^{d}(L_i(t))^2 \leq \sum_{i=0}^{d} \frac{2}{\pi\sqrt{1-t^2}} = \frac{2(d+1)}{\pi\sqrt{1-t^2}} = 2v(t)$$

That is, the leverage function is upper bounded by the Chebyshev measure, which intuitively implies that $O(d\log d)$ samples from the Chebyshev measure suffice to recover a polynomial for L_2 regression. Formally, for L_2 regression, this technique can be analyzed using the tools in [CP19a] or [RW12].

4.2 Bounding the Lewis Weights for $p = 1$ Having covered the L_2 case, we now focus on $p = 1$, where the leverage function is no longer sufficient. We turn to Lewis weights, and start by considering the standard matrix setting:

DEFINITION 4.3. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, and $p \geq 0$. Then the ℓ_p Lewis weights for $\boldsymbol{A}$ are the unique weights $w_p[\boldsymbol{A}](1), \ldots, w_p[\boldsymbol{A}](n)$ such that*

$$\tau\left[\overline{\boldsymbol{W}}^{\frac{1}{2}-\frac{1}{p}}\boldsymbol{A}\right](i) = w_p[\boldsymbol{A}](i)$$

for all $i \in [n]$, where $\overline{\boldsymbol{W}} \in \mathbb{R}^{n\times n}$ is the corresponding diagonal matrix with $\overline{\boldsymbol{W}}_{i,i} = w_p[\boldsymbol{A}](i)$.

[CP15] show several important properties of Lewis weights:

1. When $p \in [1,2]$, sampling $O(d\log d)$ rows of $\boldsymbol{A}$ with respect to its Lewis weights suffice to recover an ℓ_p subspace embedding.

2. If some other weights $w_1, \ldots, w_n$ have $\frac{1}{C} \leq \frac{\tau[\boldsymbol{W}^{\frac{1}{2}-\frac{1}{p}}\boldsymbol{A}](i)}{w_i} \leq C$ for all $i \in [n]$ and some constant C, where $\boldsymbol{W}_{i,i} = w_i$, then $w_1, \ldots, w_n$ are close to the true Lewis weights.

In particular, if we can find any such w's, then we can sample $O(d\log d)$ rows of $\boldsymbol{A}$ with respect to $w_1, \ldots, w_n$ and still get an ℓ_p subspace embedding, which suffices to recover a near-optimal solution to ℓ_p regression. This motivates our approach, where we show that the Chebyshev Measure $v(t)$ nearly satisfies this guarantee.

We start by defining Lewis weights for operators:

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

DEFINITION 4.4. *For an operator $\mathcal{A} : \mathbb{R}^{d+1} \to L_1([-1,1])$, a Lewis weight function for $\mathcal{A}$ satisfies*

$$w_p[\mathcal{A}](t) = \tau[\overline{\mathcal{W}}^{\frac{1}{2}-\frac{1}{p}}\mathcal{A}](t)$$

for all $t \in [-1,1]$, where $\overline{\mathcal{W}}$ is the corresponding diagonal operator such that $[\overline{\mathcal{W}}x](t) = w_p[\mathcal{A}](t) \cdot x(t)$ for any function x.

The Chebyshev Measure will not satisfy this strict equality criteria, so we instead consider the approximate criteria:

DEFINITION 4.5. *For an operator $\mathcal{A} : \mathbb{R}^{d+1} \to L_1([-1,1])$, a function $w(t)$ is a $C-$*Almost Lewis Weight Function *for $\mathcal{A}$ if*

$$\frac{1}{C} \leq \frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{A}](t)}{w(t)} \leq C$$

for all $t \in [-1,1]$, where $\mathcal{W}$ is the corresponding diagonal operator such that $[\mathcal{W}x](t) = w(t) \cdot x(t)$ for any function x. We often refer to $\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{A}](t)}{w(t)}$ as the Lewis Weight Fixpoint Ratio.

Similarly to the L_2 case, we relate the leverage function to a class of orthogonal polynomials. However, for $p \neq 2$, the Legendre polynomials do not make the columns of $\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{A}$ orthogonal. For $p = 1$, we turn to Chebyshev Polynomials of the Second Kind, denoted $U_i(t)$, which satisfy $\int_{-1}^{1} U_i(t)U_j(t)\sqrt{1-t^2}dt = \frac{\pi}{2}\mathbb{1}_{[i=j]}$.

THEOREM 4.1. *Let $v(t) := \frac{d+1}{\pi\sqrt{1-t^2}}$, $\mathcal{V}$ be the diagonal operator for $v(t)$, and $U_i(t)$ be the degree i Chebyshev polynomial of the second kind. Then,*

$$\frac{\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t)}{v(t)} = 1 + \frac{1 - U_{2(d+1)}(t)}{2(d+1)}$$

Proof. Let $\boldsymbol{U}$ be the change-of-basis matrix such that $\mathcal{P}\boldsymbol{U}$ has columns that are Chebyshev polynomials of the second kind. We first verify the orthogonality by simplifying the denominator of $\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t) = \max_{\mathbf{x}} \frac{([\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\boldsymbol{U}\mathbf{x}](t))^2}{\|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\boldsymbol{U}\mathbf{x}\|_2^2}$:

$$\begin{aligned}
\|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\boldsymbol{U}\mathbf{x}\|_2^2 &= \int_{-1}^{1} \left(\sum_{i=0}^{d} x_i U_i(s) \frac{1}{\sqrt{v(s)}}\right)^2 ds \\
&= \frac{\pi}{d+1} \sum_{i=0}^{d}\sum_{j=0}^{d} x_i x_j \int_{-1}^{1} U_i(s)U_j(s)\sqrt{1-s^2}ds \\
&= \frac{\pi^2}{2(d+1)}\|\mathbf{x}\|_2^2
\end{aligned}$$

With this orthogonality, we can rewrite the rescaled leverage scores as a squared row-norm:

$$\begin{aligned}
\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t) &= \max_{\mathbf{x}\in\mathbb{R}^{d+1}} \frac{\frac{1}{v(t)}([\mathcal{P}\boldsymbol{U}\mathbf{x}](t))^2}{\|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\boldsymbol{U}\mathbf{x}\|_2^2} \\
&= \frac{2(d+1)}{\pi^2 v(t)} \max_{\mathbf{x}\in\mathbb{R}^{d+1}} \frac{([\mathcal{P}\boldsymbol{U}\mathbf{x}](t))^2}{\|\mathbf{x}\|_2^2} \\
&= \frac{2(d+1)}{\pi^2 v(t)} \max_{\|\mathbf{x}\|_2=1} \left(\sum_{i=0}^{d} x_i U_i(t)\right)^2 \\
&= \frac{2(d+1)}{\pi^2 v(t)} \sum_{i=0}^{d} (U_i(t))^2
\end{aligned}$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

We now simplify this sum-of-squares term by using the specialized trigonometric structure of the Chebyshev polynomials of the second kind. Letting $\theta := \cos(t)$, note that $U_i(t) = \frac{\sin((i+1)\theta)}{\sqrt{1-t^2}}$ and the Chebyshev polynomials of the *first* kind have $T_i(t) = \cos(i\theta)$. Then,

$$(U_i(t))^2 = \frac{\sin^2((i+1)\theta)}{1-t^2} = \frac{\frac{1}{2} - \frac{1}{2}\cos(2(i+1)\theta)}{1-t^2} = \frac{\frac{1}{2} - \frac{1}{2}T_{2(i+1)}(\theta)}{1-t^2} = \frac{1}{2(1-t^2)} \cdot (1 - T_{2(i+1)}(t))$$

so that $\sum_{i=0}^{d}(U_i(t))^2 = \frac{1}{2(1-t^2)}\left((d+1) - \sum_{i=0}^{d} T_{2(i+1)}(t)\right)$. Using the relation $U_k(t) = 2\sum_{\text{even } j=1}^{k} T_j(t) - 1$ for even k, we simplify this summation as $\sum_{i=0}^{d} T_{2(i+1)}(t) = \frac{1}{2}U_{2(d+1)}(t) + \frac{1}{2} - T_0(t)$. Since $T_0(t) = 1$, $\sum_{i=0}^{d} T_{2(i+1)}(t) = \frac{1}{2}U_{2(d+1)}(t) - \frac{1}{2}$. Returning to the rescaled leverage function,

$$\begin{aligned}
\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t) &= \frac{2(d+1)}{\pi^2 v(t)} \sum_{i=0}^{d} (U_i(t))^2 \\
&= \frac{2(d+1)}{\pi^2 v(t)} \cdot \frac{d+1}{2(1-t^2)} \left(1 + \frac{1 - U_{2(d+1)}(t)}{2(d+1)}\right) \\
&= v(t) \left(1 + \frac{1 - U_{2(d+1)}(t)}{2(d+1)}\right),
\end{aligned}$$

which completes the proof. ∎

Recall that for $v(t)$ to be almost Lewis weights for $\mathcal{P}$, we need $\frac{\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t)}{v(t)} = \Theta(1)$ for all $t \in [-1, 1]$. Since $\frac{-1}{\sqrt{1-t^2}} \le U_i(t) \le \frac{1}{\sqrt{1-t^2}}$, we can see that Theorem 4.1 satisfies this criteria for almost all t:

COROLLARY 4.1. *For $|t| = 1 - O(\frac{1}{d^2})$, we have $\frac{1}{\alpha} \le \frac{\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t)}{v(t)} \le \alpha$ for some constant α.*

We prove this formally in Section 7.1.2. For $|t| \to 1$, we know that $\left|U_{2(d+1)}(t)\right| \to 2(d+1)$, so that $\frac{\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t)}{v(t)} \to 0$, meaning that the almost Lewis weight property does not hold. So, while the Chebyshev measure seems to match the Lewis weights for most t, it is wrong for t close to the "endcaps" at -1 and 1.

To understand why the Chebyshev measure fails at the endcaps, we note an important property of the leverage function. By the Markov Brother's Inequality, the leverage function is at most $O(d^2)$ for all $t \in [-1, 1]$. However, the Chebyshev measure is unbounded as $|t| \to 1$. So, there must be a gap between these two distributions.

To resolve this gap, we analyze the *Clipped Chebyshev Measure* $w(t)$, shown in Figure 4, which lies below the true Chebyshev measure $v(t)$, and which only differs in this endcap region:

DEFINITION 4.6. *The Clipped Chebyshev Measure is the function* $w(t) := \min\{C(d+1)^2, \frac{(d+1)}{\pi\sqrt{1-t^2}}\}$.

With a more involved analysis relegated to Section 7, we show that 1) $\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t) = \tilde{\Theta}(d^2)$ in the endcaps and 2) $\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t) = \Theta(\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t))$ for $|t| \le 1 - O(\frac{1}{d^2})$. This final step completes our first major technical claim:

LEMMA 4.1. (THEOREM 2.2 FOR $p = 1$) *There are fixed constants c_1, c_2 such that, for $p = 1$ and $t \in [-1, 1]$,*

$$\frac{c_1}{\log^3 d} \le \frac{\tau[\mathcal{W}^{1/2-1/p}\mathcal{P}](t)}{w(t)} \le c_2.$$

The full proof of Theorem 2.2 for general $p \in [\frac{2}{3}, 2]$, is discussed next, in Section 4.3.

4.3 Bounding the Lewis Weights for $p \in (\frac{2}{3}, 2)$ To generalize the Lewis weight analysis for $p = 1$, we find a different orthogonal polynomial that nearly achieves the $C-$almost Lewis weight property. We turn to Jacobi Polynomials:

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

DEFINITION 4.7. *The normalized Jacobi Polynomial of degree d with parameters α and β, denoted $J_d^{(\alpha,\beta)}$, defines the polynomials orthogonal with $\int_{-1}^{1} J_i^{(\alpha,\beta)}(t)J_j^{(\alpha,\beta)}(t)(1-t)^\alpha(1+t)^\beta = \mathbb{1}_{[i=j]}$.*

In particular, we look at the subclass of *Gegenbauer/Ultraspherical polynomials* which have $\alpha = \beta$, so we use the truncated notation $J_d^{(\alpha)}$ and note they are orthogonal with$\int_{-1}^{1} J_i^{(\alpha)}(t)J_j^{(\alpha)}(t)(1-t^2)^\alpha = \mathbb{1}_{[i=j]}$. Note that Legendre polynomials coincide with $\alpha = 0$, while Chebyshev polynomial of the second kind coincide with $\alpha = \frac{1}{2}$, so this class of polynomials certainly interpolates between the $p = 1$ and $p = 2$ orthogonal polynomials. We now show that Gegenbauer polynomials are the correct orthogonal polynomial for L_p Lewis weights:

THEOREM 4.2. *For all $p \in [\frac{2}{3}, 2]$ and $|t| \le 1 - O(\frac{1}{d^2})$, we have $\frac{1}{C_0} \le \frac{\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{v(t)} \le C_0$ for some universal constant C_0.*

Proof. We first show that fixing $\alpha = \frac{1}{p} - \frac{1}{2}$ and letting $\boldsymbol{U}$ be the corresponding change-of-basis matrix makes $\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\boldsymbol{U}$ have orthogonal columns:

$$\begin{aligned}
\|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\boldsymbol{U}\mathbf{x}\|_2^2 &= \int_{-1}^{1}\left(\left(\frac{d+1}{\pi\sqrt{1-s^2}}\right)^{\frac{1}{2}-\frac{1}{p}}\sum_{i=0}^{d} x_i J_i^{(\alpha)}(s)\right)^2 ds \\
&= \left(\frac{d+1}{\pi}\right)^{1-\frac{2}{p}}\int_{-1}^{1}\left(\sum_{i=0}^{d} x_i J_i^{(\alpha)}(s)\left((1-s^2)^{-\frac{1}{2}}\right)^{\frac{1}{2}-\frac{1}{p}}\right)^2 ds \\
&= \left(\frac{d+1}{\pi}\right)^{1-\frac{2}{p}}\sum_{i=0}^{d}\sum_{j=0}^{d} x_i x_j \int_{-1}^{1} J_i^{(\alpha)}(s)\ J_j^{(\alpha)}(s)\ (1-s^2)^{(\frac{1}{p}-\frac{1}{2})} ds \\
&= \left(\frac{d+1}{\pi}\right)^{1-\frac{2}{p}}\sum_{i=0}^{d}\sum_{j=0}^{d} x_i x_j \mathbb{1}_{[i=j]} \\
&= \left(\frac{d+1}{\pi}\right)^{1-\frac{2}{p}}\|\mathbf{x}\|_2^2
\end{aligned}$$

and so we can reduce $\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)$ to a squared row-norm:

$$\begin{aligned}
\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) &= \max_{\mathbf{x}\in\mathbb{R}^{d+1}} \frac{([\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2} \\
&= (\tfrac{\pi}{d+1})^{1-\frac{2}{p}} \max_{\|\mathbf{x}\|_2=1} ([\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2 \\
&= (\tfrac{\pi}{d+1})^{1-\frac{2}{p}} (\tfrac{d+1}{\pi\sqrt{1-t^2}})^{1-\frac{2}{p}} \max_{\|\mathbf{x}\|_2=1} ([\mathcal{P}\mathbf{x}](t))^2 \\
&= (1-t^2)^{-(\frac{1}{2}-\frac{1}{p})}\sum_{i=0}^{d}(J_i^{(\alpha)}(t))^2
\end{aligned}$$

Unlike the $p = 1$ case, we are not aware of any way to simplify this sum of squares exactly, so we instead provide nearly matching upper and lower bounds. For the upper bound, Theorem 1 from [NEM94] says that

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

$(J_i^{(\alpha)}(t))^2 \le \frac{C_\alpha}{\pi} \cdot (1-t^2)^{-(\alpha+\frac{1}{2})}$. We then bound

$$\begin{aligned}
\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) &\le (1-t^2)^{-(\frac{1}{2}-\frac{1}{p})} \sum_{i=0}^{d} \frac{C_\alpha}{\pi}(1-t^2)^{-(\alpha+\frac{1}{2})} \\
&= (1-t^2)^{-(\frac{1}{2}-\frac{1}{p})} \sum_{i=0}^{d} \frac{C_\alpha}{\pi}(1-t^2)^{-\frac{1}{p}} \\
&= (1-t^2)^{-\frac{1}{2}}(d+1)\frac{C_\alpha}{\pi} \\
&= C_\alpha \frac{d+1}{\pi\sqrt{1-t^2}} \\
&= C_\alpha v(t)
\end{aligned}$$

To achieve the lower bound, we appeal to a different form of an orthogonal polynomial guarantee. We rephrase $\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)$ in terms of the Generalized Christoffel Function $\lambda_d(z,2,t) := \min_{q:\deg(q)\le d} \frac{\int_{-1}^{1}(q(s))^2 z(s)ds}{(q(t))^2}$, where $z(s) := (1-s^2)^{\frac{1}{p}-\frac{1}{2}}$, as defined in Equation 1.5 of [EN92].

$$\begin{aligned}
\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) &= \max_{\mathbf{x}\in\mathbb{R}^{d+1}} \frac{([\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2} \\
&= (\tfrac{\pi}{d+1})^{1-\frac{2}{p}}(v(t))^{1-\frac{2}{p}} \max_{q:\deg(q)\le d} \frac{(q(t))^2}{\int_{-1}^{1}(q(s))^2 z(s)ds} \\
&= (\tfrac{\pi}{d+1})^{1-\frac{2}{p}}(\tfrac{d+1}{\pi})^{1-\frac{2}{p}}(1-t^2)^{\frac{-1}{2}(1-\frac{2}{p})} \max_{q:\deg(q)\le d} \frac{(q(t))^2}{\int_{-1}^{1}(q(s))^2 z(s)ds} \\
&= (1-t^2)^{\frac{1}{p}-\frac{1}{2}} \frac{1}{\min_{q:\deg(q)\le d} \frac{\int_{-1}^{1}(q(s))^2 z(s)ds}{(q(t))^2}} \\
&= (1-t^2)^{\frac{1}{p}-\frac{1}{2}} \frac{1}{\lambda_d(z,2,t)}
\end{aligned}$$

In Appendix E.1 we adapt Theorem 2.1 of [EN92] to show that $\lambda_d(z,2,t) \le \frac{C}{d-1}(1-t^2)^{\frac{1}{p}}$ for some universal constant C when $|t| \le 1 - O(\frac{1}{d^2})$. With this bound, we get $\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) \ge (1-t^2)^{-\frac{1}{2}}\frac{d-1}{C}$, so we can show the lower bound required by the almost Lewis weight property:

$$\frac{\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{v(t)} \ge \frac{(1-t^2)^{-\frac{1}{2}}\frac{d-1}{C}}{(1-t^2)^{-\frac{1}{2}}\frac{d+1}{\pi}} = \frac{\pi(d-1)}{C(d+1)} \ge \frac{\pi}{3C}$$

And so, we find that $\frac{\pi}{3C} \le \frac{\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{v(t)} \le C_\alpha$, completing the proof. ∎

Again, we see that the Chebyshev measure satisfies the almost Lewis weight property for most $t \in [-1,1]$, but this does not work in the endcaps. To remedy this issue, we again appeal to the clipped Chebyshev measure, resulting in Theorem 2.2

THEOREM 2.2 RESTATED. *There are universal constants c_1, c_2 such that, for all $p \in [\frac{2}{3}, 2]$ and $t \in [-1,1]$,*

$$\frac{c_1}{\log^3 d} \le \frac{\tau[\mathcal{W}^{1/2-1/p}\mathcal{P}](t)}{w(t)} \le c_2.$$

The full proof using this clipped measure is deferred to Section 7.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

4.4 Constant-Factor Approximation In order to achieve a constant-factor approximation to the L_p polynomial regression problem, we want to use Theorem 2.2 to create a subspace embedding guarantee. However, as discussed in Section 2.2, Lewis weight guarantees have a logarithmic dependence on the number of rows of the full matrix, which is infinite for $\mathcal{P}$.

Beyond Lewis weight sampling, it is known that matrix L_p sensitivity sampling can be done with a suboptimal dependence on the dimension d, but without any dependence on the number m of rows within the analysis. So, we bound the L_p sensitivity function of $\mathcal{P}$, showing that $\tilde{O}(d^5)$ samples drawn *uniformly* from $[-1,1]$ creates a subspace embedding from the $\mathcal{P}$ operator to a tall-and-skinny matrix $\boldsymbol{A}$. With this sensitivity result, we can solve the problem in Theorem 1.1 with $\tilde{O}(d^5)$ samples:

DEFINITION 4.8. (L_p SENSITIVITY FUNCTION) *For an operator $\mathcal{A} : \mathbb{R}^{d+1} \to L_p([-1,1])$, the L_p sensitivity function for $\mathcal{A}$ at time $t \in [-1,1]$ is*

$$\psi_p[\mathcal{A}](t) := \max_{\mathbf{x}\in\mathbb{R}^{d+1}} \frac{|[\mathcal{A}\mathbf{x}](t)|^p}{\|\mathcal{A}\mathbf{x}\|_p^p}.$$

We show that the sensitivities of L_p regression are bounded.

LEMMA 4.2. (UNIFORM SENSITIVITY BOUND) *For all $t \in [-1,1]$ and $p \geq 1$, we have $\psi_p[\mathcal{P}](t) \leq d^2(p+1)$*

Proof. Note that $\psi_p[\mathcal{P}] := \max_{\mathbf{x}\in\mathbb{R}^{d+1}} \frac{|[\mathcal{P}\mathbf{x}](t)|^p}{\|\mathcal{P}\mathbf{x}\|_p^p} = \max_{q:\deg(q)\leq d} \frac{|q(t)|^p}{\int_{-1}^1 |q(s)|^p ds}$ Without loss of generality we take $q(t) = 1$. Let $C := \max_{s\in[-1,1]} |q(x)|$ and $s^* := \operatorname{argmax}_{s\in[-1,1]} |q(x)|$. By the Markov brothers' inequality, we have $|q(s^*+s)| \geq C - Cd^2 s \geq 0$ for any $|s| \leq \frac{1}{d^2}$. Then we can lower bound the integral in the denominator of ψ_p by

$$\int_{-1}^1 |q(s)|^p \, ds \geq \int_0^{\frac{1}{d^2}} (C - Cd^2 s)^p \, ds = \frac{-1}{Cd^2(p+1)}(C - Cd^2 x)^{p+1}\Big|_0^{1/d^2} \geq \frac{1}{d^2(p+1)}$$

so that

$$\psi_p[\mathcal{P}](t) = \frac{|q(t)|^p}{\int_{-1}^1 |q(x)|^p \, dx} \leq d^2(p+1)$$

∎

Next we show that since uniform sampling is oversampling with respect to the sensitivities, we can get an L_p subspace embedding with $\tilde{O}(d^5)$ samples:

THEOREM 4.3. *Let $p \geq 1$ and suppose $s_1, \ldots, s_{n_0}$ are drawn uniformly from $[-1,1]$. Let $\boldsymbol{A} \in \mathbb{R}^{n_0\times(d+1)}$ be the associated Vandermonde matrix, so that $\boldsymbol{A}_{i,j} = s_i^{j-1}$. Let $\mathbf{b} \in \mathbb{R}^{n_0}$ be the evaluations of f, so that $\mathbf{b}(i) = f(s_i)$. For $n_0 = O\left(d^5 p^2 2^p \log d\right)$, there exists a universal constant c such that the sketched solution $\mathbf{x}_c = \operatorname{argmin}_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p$ satisfies*

$$\|\mathcal{P}\mathbf{x}_c - f\|_p \leq c \min_{\mathbf{x}\in\mathbb{R}^{d+1}} \|\mathcal{P}\mathbf{x} - f\|_p$$

with probability at least $\frac{11}{12}$.
Further, let $\varepsilon \in (0,1)$ and suppose $\|f\|_p \leq C \min_{\mathbf{x}} \|\mathcal{P}\mathbf{x} - f\|_p$. If $n_0 = O\left(\frac{1}{\varepsilon^{O(p^2)}} d^5 p^{O(p)} \log \frac{d}{\varepsilon}\right)$, then

$$\|\mathcal{P}\hat{\mathbf{x}} - f\|_p^p \leq (1+\varepsilon) \min_{\mathbf{x}} \|\mathcal{P}\mathbf{x} - f\|_p^p$$

with probability at least $\frac{11}{12}$. In particular, suppose $\mathbf{x}_c$ is computed from sampling f uniformly at least $O(d^5 p^2 2^p \log(d))$ times, we let $\hat{f}(t) := f(t) - [\mathcal{P}\mathbf{x}_c](t)$, and compute $\hat{\mathbf{x}}$ by sampling $\hat{f}$ uniformly at least $O\left(\frac{1}{\varepsilon^{O(p^2)}} d^5 p^{O(p)} \log \frac{d}{\varepsilon}\right)$ times. Then, if we let $\tilde{\mathbf{x}} := \mathbf{x}_c + \hat{\mathbf{x}}$, we have

$$\|\mathcal{P}\tilde{\mathbf{x}} - f\|_p^p \leq (1+\varepsilon) \min_{\mathbf{x}\in\mathbb{R}^{d+1}} \|\mathcal{P}\mathbf{x} - f\|_p^p$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

The proof of this theorem is a standard sensitivity sampling analysis combined with our bounds on the L_p sensitivities, so it is deferred to Appendix A.

To decrease this sample complexity further, we apply Lewis weight subsampling to the matrix $\boldsymbol{A}$. Since the rows of $\boldsymbol{A}$ are drawn *uniformly* from $[-1, 1]$, we can show that the Lewis weights of $\boldsymbol{A}$ closely approximate the Lewis weights of $\mathcal{P}$. So, by Theorem 2.2, we know that the Chebyshev measure upper bounds the Lewis weights of $\boldsymbol{A}$. That is, *we can bound the Lewis weights of $\boldsymbol{A}$ without ever even building the matrix.* Formally, we give the following guarantee:

THEOREM 4.4. *Let $\boldsymbol{A}$, and n_0 as in either part of Theorem 4.3. Then, with probability $\frac{11}{12}$, for all $i \in [n_0]$, the ℓ_p Lewis weight of $\boldsymbol{A}$ at row i is at most $\frac{1}{n_0} v(s_i) \operatorname{polylog}(d)$ and at least $\frac{1}{n_0 \operatorname{polylog}(d)} w(s_i)$.*

Proof. Let $\boldsymbol{W} \in \mathbb{R}^{n_0 \times n_0}$ be a diagonal matrix that represents our candidate ℓ_p Lewis weights for $\boldsymbol{A}$, with $\boldsymbol{W}_{ii} := \gamma w(s_i)$, where $\gamma := \frac{2}{n_0}$ is a rescaling factor. In Appendix B we use a standard ε-net argument to show the spectral approximation

$$\frac{1}{2}\mathcal{P}^\top \mathcal{W}^{1-\frac{2}{p}} \mathcal{P} \preceq \gamma^{-\frac{2}{p}} \boldsymbol{A}^\top \boldsymbol{W}^{1-\frac{2}{p}} \boldsymbol{A} \preceq 2\mathcal{P}^\top \mathcal{W}^{1-\frac{2}{p}} \mathcal{P}$$

holds with probability $\frac{11}{12}$. We condition on this event.

Then note the inner-product form of the leverage scores: $\tau[\boldsymbol{A}]_i = \mathbf{a}_i^\top (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \mathbf{a}_i$ where $\mathbf{a}_i$ is the i^{th} row of $\boldsymbol{A}$, and $\tau[\mathcal{P}](t) = \mathbf{p}_t^\top (\mathcal{P}^\top \mathcal{P})^{-1} \mathbf{p}_t$ where $\mathbf{p}_t := [1\ t\ t^2\ \ldots\ t^d]$ is the row of $\mathcal{P}$ at time t (Theorem 5 from [AKM+19] or Lemma 1 from [Mey22]). Then we can examine the rescaled leverage scores:

$$\begin{aligned}
\tau[\boldsymbol{W}^{\frac{1}{2}-\frac{1}{p}} \boldsymbol{A}](i) &= (\boldsymbol{W}_{ii})^{1-\frac{2}{p}} \mathbf{a}_i^\top (\boldsymbol{A}^\top \boldsymbol{W}^{1-\frac{2}{p}} \boldsymbol{A})^{-1} \mathbf{a}_i \\
&\leq 2(\boldsymbol{W}_{ii})^{1-\frac{2}{p}} \gamma^{\frac{2}{p}} \mathbf{a}_i^\top (\mathcal{P}^\top \mathcal{W}^{1-\frac{2}{p}} \mathcal{P})^{-1} \mathbf{a}_i \\
&= 2(\gamma w(s_i))^{1-\frac{2}{p}} \gamma^{\frac{2}{p}} \mathbf{p}_{s_i}^\top (\mathcal{P}^\top \mathcal{W}^{-1} \mathcal{P})^{-1} \mathbf{p}_{s_i} \\
&= 2\gamma\ \tau[\mathcal{W}^{-1}\mathcal{P}](s_i)
\end{aligned}$$

and we can similarly show that $\tau[\boldsymbol{W}^{\frac{1}{2}-\frac{1}{p}} \boldsymbol{A}](i) \geq \frac{1}{2}\gamma \tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}} \mathcal{P}](s_i)$. So now we can use Theorem 2.2 to show the almost Lewis weight property holds on $\boldsymbol{A}$:

$$\frac{\tau[\boldsymbol{W}^{\frac{1}{2}-\frac{1}{p}} \boldsymbol{A}](i)}{\boldsymbol{W}_{ii}} \leq \frac{2\gamma\ \tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}} \mathcal{P}](s_i)}{\gamma\ w(s_i)} = 2\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}} \mathcal{P}](s_i)}{w(s_i)} \leq \log^3(d)$$

and similarly we can show the lower bound $\frac{\tau[\boldsymbol{W}^{\frac{1}{2}-\frac{1}{p}} \boldsymbol{A}](i)}{\boldsymbol{W}_{ii}} \geq \log^3(d)$. Therefore, $\boldsymbol{W}_{ii} = \frac{2}{n_0} w(t)$ are ℓ_p almost Lewis weights for $\boldsymbol{A}$. Further, since $v(t) \geq w(t)$, we have that $\frac{C}{n_0} v(t)$ upper bound the ℓ_p Lewis weights for $\boldsymbol{A}$ for some constant C. ∎

This naïvely suggests an $\tilde{O}(d^5)$ runtime algorithm to pick $O(d \operatorname{polylog} d)$ samples that give optimal L_p regression: sample $n_0 = O(d^5 \log d)$ times uniformly from $[-1, 1]$, and for each sample, throw it away with probability $1 - \min\{\frac{1}{n_0} v(s_i) \operatorname{polylog}(d), 1\}$. Then, with high probability, $O(d)$ samples remain and the resulting subsampled matrix is an L_p subspace embedding. Formally, this argument uses the following result from [CP15]:

THEOREM 4.5. (THEOREM 7.1 FROM [CP15][7]) *Let $\boldsymbol{A} \in \mathbb{R}^{n_0 \times d+1}$ and $p \in [1, 2]$. Let $w_p[\boldsymbol{A}](1), \ldots, w_p[\boldsymbol{A}](n_0)$ be the ℓ_p Lewis weights of $\boldsymbol{A}$, and let $\tilde{w}_i \geq C w_p[\boldsymbol{A}](i)$ for all i such that $\sum_i \tilde{w}_i = \tilde{O}(d)$. Define probabilities $p_i := \min\{1, \frac{m}{n_0} \tilde{w}_i\}$, and build the diagonal matrix $\boldsymbol{S} \in \mathbb{R}^{n_0 \times n_0}$ such that $\boldsymbol{S}_{ii}$ takes value $\frac{1}{(p_i)^{1/p}}$ with probability p_i and is 0 otherwise. Remove the rows of $\boldsymbol{S}$ that are all zero. Suppose we pick m, the expected number of remaining rows, to be $m = O(d \operatorname{polylog}(d))$. Then with probability $\frac{11}{12}$, for all $\mathbf{x} \in \mathbb{R}^{d+1}$, we have $\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}\|_p \approx_2 \|\boldsymbol{A}\mathbf{x}\|_p$.*

This $\tilde{O}(d^5)$ time algorithm certainly suffices to give the near-optimal sample complexity for constant ε, but we can improve the time complexity. In particular, since we exactly know the distribution of $s_1, \ldots, s_{n_0}$ and the probabilities of the coins $p_1, \ldots, p_{n_0}$, we can directly compute the marginal distribution of times that result from both sampling procedures:

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

LEMMA 4.3. *Suppose n_0 time samples are drawn uniformly from $[-1,1]$, and each sample is thrown away with probability $1-\min\{\frac{m}{n_0}\frac{1}{\sqrt{1-s_i^2}},1\}$. Let n denote the number of remaining samples. Then n is distributed as $B(n_0,O(\frac{m}{n_0}))$, and with probability $\frac{99}{100}$ the resulting samples cannot be distinguished from iid samples from the Chebyshev measure.*

This short lemma is proven in Appendix C. Taking $n_0 = O(d^5 \operatorname{polylog} d)$ and $m = O(d \operatorname{polylog} d)$, we get $n \sim B(n_0, 1/\tilde{O}(d^4))$ so that $n = d \operatorname{polylog} d$ with very high probability. So, this lemma tells us that instead of sampling $\tilde{O}(d^5)$ times uniformly, we can just sample d polylog(d) samples from the Chebyshev distribution. In summary, we arrive at the following:

COROLLARY 4.2. *Let $\boldsymbol{A}$, and n_0 as in either part of Theorem 4.3. Let $m = O(d \operatorname{polylog} d)$. Suppose an algorithm samples $n \sim B(n_0, O(\frac{m}{n_0}))$ and runs Algorithm 1. Then, the matrix $\boldsymbol{SA}$ on line 4 of the algorithm is a subspace embedding for $\mathcal{P}$: $\frac{1}{C}\|\mathcal{P}\mathbf{x}\|_p^p \le \|\boldsymbol{SA}\mathbf{x}\|_p^p \le C\|\mathcal{P}\mathbf{x}\|_p^p$ for all $\mathbf{x} \in \mathbb{R}^{d+1}$.*

We now state the overall correctness of the algorithm for constant factor approximation for $p \ge 1$:

THEOREM 4.6. *Let $p \ge 1$ and $n_0 = O\left(d^5 p^2 2^p \log d\right)$. Suppose an algorithm samples $n \sim B(n_0, 1/\tilde{O}(d^4))$ and runs Algorithm 1. Then, with probability $\frac{2}{3}$, the resulting polynomial $\hat{q}$ satisfies*

$$\|\hat{q} - f\|_p^p \le O(1) \min_{q:deg(q)\le d} \|q - f\|_p^p$$

The correctness of this theorem follows from combining Corollary 4.2 with Lemma A.1 from [MMM+22], which says that unbiased subspace embedding suffices for constant-factor error in regression. While there is randomness in the sample complexity, we have that with very high probability $n = O(d \operatorname{polylog} d)$. Finally, we emphasize that Theorem 4.6 holds for all $p \ge 1$ due to the result from [MMM+22]. Thus we will ultimately also use this algorithm as a subroutine for L_p polynomial regression for $p \ge 2$.

4.5 $(1+\varepsilon)$-Approximation Given the constant factor approximation in the previous section, we can now build an algorithm that outputs a $(1+\varepsilon)$-approximation for the L_p regression problem when $p \in [1,2]$. First, we recall an algorithm from [MMWY22] that samples $d \operatorname{poly}(\log d, \frac{1}{\varepsilon})$ rows of a matrix by almost-Lewis weights, reads the corresponding coordinates in the measurement vector $\mathbf{b}$, and solves the subsampled ℓ_p matrix regression problem twice, giving a $(1+\varepsilon)$ error solution. Since we know that the Chebyshev density describes the almost-Lewis weights of $\boldsymbol{A}$, we can directly appeal to this result. In particular, they prove that Algorithm 4 gives the following guarantee:

THEOREM 4.7. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times d+1}$, $\mathbf{b} \in \mathbb{R}^m$, and $p \ge 1$. Then, with probability 0.98, Algorithm 4 with $n = O(d^{\max(1,p/2)} \frac{\log^2(d)\log(m)}{\varepsilon^{\min(2p+5,p+7)}})$ returns a vector $\tilde{\mathbf{x}} \in \mathbb{R}^{d+1}$ such that $\|\boldsymbol{A}\tilde{\mathbf{x}} - \mathbf{b}\|_p \le (1+\varepsilon)\min_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p$.*

We remark that although Theorem 4.7 matches the guarantee given by Theorem 3.4 in [MMWY22][8], Algorithm 4 does not quite match the corresponding Algorithm 2 given by [MMWY22]. Observe that each row is sampled without replacement with probability proportional to its Lewis weight in Algorithm 3, whereas a fixed number of rows are sampled by [MMWY22], so that each row is sampled with replacement with probability proportional to its Lewis weight. However, the correctness of Algorithm 3 follows from the analysis of Theorem 3.4 in [MMWY22] by zooming into Claim 3.14 and just using a sampling matrix $\boldsymbol{S}$ defined by without-replacement sampling instead of the with-replacement matrix used. None of the concentrations actually change at the end of the day. We show an example of such a Bernstein bound later in this paper, in the proof of Lemma 5.7.

Overall, Theorem 4.7 show that Algorithm 4 finds a near-optimal solution to the uniform-sampled problem for $p \in [1,2]$. By the reduction from two-stage to one-stage sampling, this then implies that Algorithm 2 finds a near-optimal solution to the L_p polynomial regression problem. So, we have now proven our L_p polynomial approximation guarantee for $p \in [1,2]$:

[8]This is following the first version of [MMWY22] uploaded to arXiv, which uses an analysis which makes especially simple to see how Bernstein suffices for either sampling scheme.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Algorithm 3 Constant factor active ℓ_p matrix regression

Input: Vandermonde matrix $\boldsymbol{A} \in \mathbb{R}^{n_0 \times d+1}$, response vector $\mathbf{b} \in \mathbb{R}^{n_0}$, target number of samples m
Output: Approximate solution $\hat{\mathbf{x}} \in \mathbb{R}^{d+1}$ to $\min_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p$
1: Let $p_i = \min\{1, \frac{m}{n_0} \frac{1}{\sqrt{1-s_i^2}}\}$ where $s_i \in [-1, 1]$ is the time associated with row i of $\boldsymbol{A}$
2: Let $\boldsymbol{S} \in \mathbb{R}^{n_0 \times n_0}$ be a diagonal matrix with $S_{ii} = \frac{1}{(p_i)^{1/p}}$ with probability p_i, and $S_{ii} = 0$ otherwise
3: **return** $\hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}} \|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\mathbf{b}\|_p$

Algorithm 4 Relative error active ℓ_p matrix regression

Input: Matrix $\boldsymbol{A} \in \mathbb{R}^{n_0 \times d+1}$, response vector $\mathbf{b} \in \mathbb{R}^{n_0}$, target number of samples n
Output: Approximate solution $\tilde{\mathbf{x}} \in \mathbb{R}^{d+1}$ to $\min_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p$
1: Run Algorithm 3 on vector $\mathbf{b}$ with $\frac{n}{2}$ samples to get vector $\mathbf{x}_c$
2: Let $\mathbf{z} := \mathbf{b} - \boldsymbol{A}\hat{\mathbf{x}}_c$
3: Run Algorithm 3 on vector $\mathbf{z}$ with $\frac{n}{2}$ samples to get vector $\hat{\mathbf{x}}$
4: **return** $\tilde{\mathbf{x}} = \mathbf{x}_c + \hat{\mathbf{x}}$

THEOREM 1.1 RESTATED. *[For $1 \le p \le 2$] For any degree d, $p \in [1, 2]$, and accuracy parameter $\varepsilon \in (0, 1)$, there is an algorithm that queries f at $n = O(\frac{d}{\varepsilon^{2p+5}} \operatorname{polylog}(\frac{d}{\varepsilon}))$ points $t_1, \ldots, t_n$, each selected independently at random according to the Chebyshev density on $[-1, 1]$, and outputs a degree d polynomial $\hat{q}(t)$ such that, with probability at least 0.9,*

$$\|\hat{q}(t) - f(t)\|_p^p \le (1 + \varepsilon) \cdot \min_{q:\deg(q) \le d} \|q(t) - f(t)\|_p^p.$$

5 Active L_p Regression for $p > 2$

In this section, we analyze L_p regression for $p > 2$. Our analysis differs significantly from the case of $p \in [1, 2]$. In particular, while we still analyze sampling by the Chebyshev measure, in contrast to $p \in [1, 2]$, we are not able to argue that the measure approximates the L_p Lewis weights of the polynomial operate $\mathcal{P}$. Moreover, even if we could bound them, sampling by L_p Lewis weights requires $O(d^{p/2})$ samples in the worst case to approximate a p-norm regression problem [MMWY22]. There are matrices which require this rate, so to get sample complexity linear in d, we will leverage special structure of polynomials that lets us avoid these worst-case instances.

We start with a simple but useful observation from [MMM+22]. Ssuppose $f(t)$ is a polynomial of degree d, and let $r \approx p$ be an integer with $q := \frac{p}{r} \in [1, 2]$. Then, we know that $t \mapsto (f(t))^r$ is a degree rd polynomial. Since $\boldsymbol{A}$ is a Vandermonde matrix, and letting $\mathbf{x}$ be the coefficient vector for f, we thus have that

$$\|\boldsymbol{A}\mathbf{x}\|_p^p = \|\boldsymbol{B}\mathbf{y}\|_q^q,$$

where $\boldsymbol{B} \in \mathbb{R}^{n_0 \times rd+1}$ is a Vandermonde matrix generated by the same time points as $\boldsymbol{A}$ but with more columns, and where $\mathbf{y}$ is the coefficient vector for the degree rd polynomial $t \mapsto (f(t))^r$. This simple observation implies that if some sampling procedure preserves the ℓ_q norm of all degree rd polynomials, then that sampling procedure also preserves the ℓ_p norm of all degree d polynomials. In other words, it suffices to use a sampling matrix $\boldsymbol{S}$ that samples rows of $\boldsymbol{B}$ with probability proportional to upper bounds on the ℓ_q Lewis weights of $\boldsymbol{B}$. By Theorem 4.4 we already know those Lewis weights are bounded by the Chebyshev measure. So Algorithm 3, which samples rows of $\boldsymbol{A}$ by the Chebyshev measure, preserves the ℓ_p norm of $\boldsymbol{A}\mathbf{x}$ for all $\mathbf{x}$ *because* it is sampling rows by the ℓ_q Lewis weights of $\boldsymbol{B}$.

This argument suffices to get prove a subspace embedding result – i.e., that the matrix $\boldsymbol{S}$ from Algorithm 3 satisfies $\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}\|_p^p \approx_C \|\boldsymbol{A}\mathbf{x}\|_p^p$. This is sufficient to get a constant-factor regression solution, and we formally work through this in Section 5.1. To achieve error $(1 + \varepsilon)$, we need a more refined analysis that builds on the first version of [MMWY22] uploaded to arXiv[9]. Our approach still reduces from the general $p > 2$ case to some $q \le 2$, but in a less direct way than described above. An edge case of our analysis requires that when $p \in (2, 3)$, we use $r = 3$ so that $q = \frac{p}{r} \in [\frac{2}{3}, 1]$. This is the case where we use the ℓ_q Lewis weight bounds for $q < 1$.

[9] While that version is available on arXiv at time of publishing, since it is unpublished, we include a (slightly shortened and corrected) copy of everything we use in Appendix D.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

5.1 Constant Factor Approximation for $p > 2$ We start by showing that running Algorithm 3 as done in line 1 of Algorithm 4 achieves a constant-factor regression guarantee. Formally, we rely on the following result from [MMM+22], where $\psi_p[\boldsymbol{A}](i) := \max_{\mathbf{x}} \frac{|[\boldsymbol{A}\mathbf{x}](i)|^p}{\|\boldsymbol{A}\mathbf{x}\|_p^p}$ is the ℓ_p sensitivity score of $\boldsymbol{A}$ at row i:

THEOREM 5.1. *[MMM+22] Given $p > 2$, let r be any integer such that $q := \frac{p}{r}$ is in $[\frac{2}{3}, 2]$. Given a Vandermonde matrix $\boldsymbol{A} \in \mathbb{R}^{n_0 \times d+1}$, let $\boldsymbol{B}$ be the Vandermonde matrix $\boldsymbol{A}$ extended to have $rd+1$ columns. Then for every vector $\mathbf{x} \in \mathbb{R}^{d+1}$, there exists a vector $\mathbf{y} \in \mathbb{R}^{rd+1}$ such that $|[\boldsymbol{A}\mathbf{x}](i)|^p = |[\boldsymbol{B}\mathbf{y}](i)|^q$. Thus if $\psi_p[\boldsymbol{A}](i)$ denotes the ℓ_p-sensitivity of the i-th row of $\boldsymbol{A}$ and $\psi_q[\boldsymbol{B}](i)$ denotes the ℓ_q-sensitivity of the i-th row of $\boldsymbol{B}$, then $\psi_p[\boldsymbol{A}](i) \leq \psi_q[\boldsymbol{B}](i)$.*

For the constant-factor approximation step, given $p > 2$, we let r be an integer such that $r \leq p < 2r$, so that $q := \frac{p}{r} \in [1, 2]$. With this choice of r, chosen such that ℓ_q is a valid norm that satisfies the triangle inequality, we will show that Algorithm 3, as run in the first line of Algorithm 4, returns a constant factor solution to $\min_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p$. Recall that $\boldsymbol{A} \in \mathbb{R}^{n_0 \times d+1}$ is a Vandermonde matrix obtained by uniformly sampling $n_0 = \text{poly}(d, p^p, \frac{1}{\varepsilon^p})$ points from $[-1, 1]$. Then let $\boldsymbol{B} \in \mathbb{R}^{n_0 \times rd+1}$ be an *expanded* Vandermonde matrix, built using the same uniform samples but with maximum degree rd. Let $d_B := rd+1$ be the number of columns in $\boldsymbol{B}$. We also let $w_q[\boldsymbol{B}](i)$ be the ℓ_q-Lewis weight of $\boldsymbol{B}$ at row i. We will analyze sampling rows of $\boldsymbol{A}$ with respect to $w_q[\boldsymbol{B}](i)$.

We first show that the sampling matrix $\boldsymbol{S}$ from Algorithm 3 is a subspace embedding:

LEMMA 5.1. *Let $\boldsymbol{A}$ and $\boldsymbol{S}$ be the matrices as in Algorithm 3. Then, with probability $\frac{99}{100}$, so long as $m = O(\frac{pd}{\varepsilon^2}\,\text{polylog}(d))$, we have that $\boldsymbol{S}$ is an ℓ_p subspace embedding:*

$$\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}\|_p^p \in (1 \pm \varepsilon)\|\boldsymbol{A}\mathbf{x}\|_p^p \qquad \forall \mathbf{x} \in \mathbb{R}^{d+1}$$

Proof. Recall Theorem 5.1, in particular that for any $\mathbf{x} \in \mathbb{R}^{d+1}$, there exists a vector $\mathbf{y} \in \mathbb{R}^{rd+1}$ such that $([\boldsymbol{B}\mathbf{y}](i))^q = ([\boldsymbol{A}\mathbf{x}](i))^p$ for all $i \in [n_0]$. We then expand the subspace embedding norm:

$$\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}\|_p^p = \sum_{i=1}^{n_0} \boldsymbol{S}_{ii}^p\,|[\boldsymbol{A}\mathbf{x}](i)|^p = \sum_{i=1}^{n_0} \frac{1}{p_i}\,|[\boldsymbol{B}\mathbf{y}](i)|^q = \|\bar{\boldsymbol{S}}\boldsymbol{B}\mathbf{y}\|_q^q$$

where $\bar{\boldsymbol{S}}_{ii} = (\boldsymbol{S}_{ii})^{p/q} = \frac{1}{(p_i)^{1/q}}$ is the sampling matrix we would use when sampling $\boldsymbol{B}$ by ℓ_q Lewis weights. So, we not only have $\|\boldsymbol{A}\mathbf{x}\|_p^p = \|\boldsymbol{B}\mathbf{y}\|_q^q$, but also have $\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}\|_p^p = \|\bar{\boldsymbol{S}}\boldsymbol{B}\mathbf{y}\|_q^q$. Then we are sampling by overestimates of the Lewis weights, since $w_q[\boldsymbol{B}](i) \leq \frac{1}{n_0}\frac{rd+1}{\sqrt{1-s_i^2}}\,\text{polylog}(d) \leq \frac{m}{n_0}\frac{1}{\sqrt{1-s_i^2}} = p_i$, which holds for $m \geq d\,\text{polylog}(d)$. So, by Theorem 4.5, we have that $\boldsymbol{S}$ is a $(1 \pm \frac{1}{2})$ ℓ_q-subspace embedding for $\boldsymbol{B}$ so long as $m = O(\frac{rd}{\varepsilon^2}\,\text{polylog}(d))$, and therefore that $\bar{\boldsymbol{S}}$ is a $(1 \pm \varepsilon)$ ℓ_p-subspace embedding for $\boldsymbol{A}$. ∎

LEMMA 5.2. *The vector $\mathbf{x}_c$ returned by line 1 of Algorithm 3 is a constant-factor solution to the overall optimization problem, with probability $\frac{99}{100}$:*

$$\|\boldsymbol{A}\mathbf{x}_c - \mathbf{b}\|_p \leq C_z \min_{\mathbf{x} \in \mathbb{R}^{d+1}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p$$

For some universal constant C_z. In particular, this implies that $\mathbf{z}$ from line 2 of Algorithm 3 has $\|\mathbf{z}\|_p \leq C_z \min_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p$.

Proof. Recall that $\mathbf{x}_c := \text{argmin}_{\mathbf{x}} \|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\mathbf{b}\|_p$, and that Lemma 5.1 shows that $\boldsymbol{S}$ is an ℓ_p subspace embedding for $\boldsymbol{A}$. Let $\mathbf{x}^* := \text{argmin}_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|$ be the true optimal regression solution. Then, by repeated use of the triangle inequality,

$$\begin{aligned}
\|\boldsymbol{A}\mathbf{x}_c - \mathbf{b}\|_p &\leq \|\boldsymbol{A}\mathbf{x}_c - \boldsymbol{A}\mathbf{x}^*\|_p + \|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p \\
&\leq 2\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}_c - \boldsymbol{S}\boldsymbol{A}\mathbf{x}^*\|_p + \|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p \\
&\leq 2(\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}_c - \boldsymbol{S}\mathbf{b}\|_p + \|\boldsymbol{S}\boldsymbol{A}\mathbf{x}^* - \boldsymbol{S}\mathbf{b}\|_p) + \|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p \\
&\leq 4\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}^* - \boldsymbol{S}\mathbf{b}\|_p + \|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p
\end{aligned}$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

where the last line follows from the optimality of $\tilde{\mathbf{x}}$. Then, since $\mathbb{E}[\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}^* - \boldsymbol{S}\mathbf{b}\|_p^p] = \|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p^p$, by Markov's inequality we bound $\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}^* - \boldsymbol{S}\mathbf{b}\|_p^p \leq 200\|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p^p$, and we conclude that

$$\|\boldsymbol{A}\mathbf{x}_c - \mathbf{b}\|_p \leq 801\|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p$$

□

5.2 Relative Error Approximation In this section, we show that the estimator $\tilde{\mathbf{x}}$ recovered on by Algorithm 4 is a $(1+\varepsilon)$-optimal estimator for $\|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p^p$. First, note we assume that $\varepsilon \leq \frac{1}{p}$ in this section, and prove that sampling $\tilde{O}(\frac{d\,2^{O(p)}}{\varepsilon^{6p+2}})$ rows suffices to recover a near-optimal estimator. If $\varepsilon > \frac{1}{p}$, then we can just run the algorithm when $\varepsilon = \frac{1}{p}$, which yields a $\tilde{O}(dp^{O(p)})$ sample complexity, so the sample complexity we promise in Theorem 1.1 suffices across all possible $\varepsilon \in (0,1)$ and $p \geq 2$.

Much of this section very closely tracks the proof of Theorem 3.4 in the first version of [MMWY22] uploaded to arXiv, with the main difference being Lemma 5.3 which uses Theorem 5.1 to define the vector $\bar{\mathbf{z}}$ with respect to the ℓ_q Lewis weights of $\boldsymbol{B}$, where the original analysis uses the ℓ_p Lewis weights of $\boldsymbol{A}$. The core of the novel analysis is used to prove Theorem 5.2. While we state and use Theorem 5.2 in this section, we do not prove it until later, in Section 5.3.

Most of this section analyzes the second call to Algorithm 3, from the line 3 of Algorithm 4. As such, we explicitly write down the notation that will be used throughout most of this section:

SETTING 5.1. *$\boldsymbol{A} \in \mathbb{R}^{n_0 \times d+1}$ is a Vandermonde matrix formed by sampling $n_0 = O(\frac{1}{\varepsilon^{O(p^2)}} d^5 p^{O(p^2)} \log \frac{d}{\varepsilon})$ times $s_1, \ldots, s_{n_0}$ uniformly at random from $[-1,1]$. r is an integer such that $\frac{1}{2}p \leq r < \frac{3}{2}p$, and $q := \frac{p}{r} \in [\frac{2}{3}, 2]$. $\boldsymbol{B} \in \mathbb{R}^{n_0 \times d_B}$ is a Vandermonde matrix formed from the same time samples $s_1, \ldots, s_{n_0}$, but with $d_B := rd+1$ columns. $w_q[\boldsymbol{B}](i)$ denotes the ℓ_q Lewis Weight of $\boldsymbol{B}$ at row i, and $\psi_p[\boldsymbol{A}](i) := \max_{\mathbf{x}} \frac{|[\boldsymbol{A}\mathbf{x}](i)|^p}{\|\boldsymbol{A}\mathbf{x}\|_p^p}$ denotes the ℓ_p sensitivity of row i of $\boldsymbol{A}$. $\mathbf{z} := \mathbf{b} - \boldsymbol{A}\mathbf{x}_c$ is the vector generated by line 2 of Algorithm 3. By Lemma 5.2, $\|\mathbf{z}\|_p \leq C_z OPT$, where $OPT = \min_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p$. $\bar{\mathbf{z}}$ is equal to $\mathbf{z}$ except that it has several entries zeroed out:*

$$\bar{\mathbf{z}}(i) := \begin{cases} \mathbf{z}(i) & |\mathbf{z}(i)| \leq \frac{OPT}{\varepsilon}(w_q[\boldsymbol{B}](i))^{1/p} \\ 0 & \text{otherwise} \end{cases}$$

Let $\boldsymbol{S} \in \mathbb{R}^{n_0 \times n_0}$ be the sample-and-rescale matrix generated in step 3 of Algorithm 4 with $m = O(\frac{d}{\varepsilon^{6.5p+2}} \operatorname{polylog}(\frac{d}{\varepsilon}))$. $C_0 := 400C_z$ is a large enough constant.

Note that r in this section might not be the same value of r taken in the constant factor analysis of Section 5.1. We explain this new choice of r in Section 5.3 in full detail, but at a high level, we will eventually want r to be odd for this analysis to go through, which will sometimes require $q \in [\frac{2}{3}, 1]$, for instance.

In the majority of this proof, we constrict ourselves to looking at vectors in the range of $\boldsymbol{A}$ which are not too much larger than OPT, defining a sort of "reasonable range of $\boldsymbol{A}$" to focus on. Rigorously, this means the upcoming lemmas will only look at vectors in the range of $\boldsymbol{A}$ with $\|\boldsymbol{A}\mathbf{x}\|_p \leq C_0 OPT$. We will eventually ensure that both $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{z}\|_p$ and $\hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}} \|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\mathbf{z}\|_p$ lie within this reasonable range.

We first examine the vector $\bar{\mathbf{z}}$ defined in Setting 5.1. Intuitively, we say that the entries of $\mathbf{z}$ that get zeroed out are so large that the reasonable range of $\boldsymbol{A}$ cannot fit them. So, we can approximate the true error by $\|\boldsymbol{A}\mathbf{x} - \mathbf{z}\|_p^p \approx \|\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}\|_p^p + \|\mathbf{z} - \bar{\mathbf{z}}\|_p^p$. That is, minimizing $\|\boldsymbol{A}\mathbf{x} - \mathbf{z}\|_p$ is effectively equivalent to minimizing $\|\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}\|_p$.

We define the zeroing-out procedure in terms of the ℓ_q Lewis weights of $\boldsymbol{B}$ here, so this is one place where we adapt the prior work to use the special structure of Vandermonde matrices. Roughly, the ℓ_p sensitivity $\psi_p[\boldsymbol{A}](i)$ measures how spikey a vector in the reasonable range can be. The Vandermonde structure lets us bound the sensitivity of $\boldsymbol{A}$ with the sensitivity of $\boldsymbol{B}$, since $\psi_p[\boldsymbol{A}](i) \leq \psi_q[\boldsymbol{B}](i)$. Then, we use the fact that all matrices have their ℓ_q sensitivities bounded by their Lewis weights for $q \leq 2$. So, we can bound the spikeyness of the i^{th} entry of a vector in the reasonable range by the ℓ_q Lewis weight of $\boldsymbol{B}$ at row i. For general matrices, the ℓ_p sensitivity $\psi_p[\boldsymbol{A}](i)$ can be $d^{\frac{p}{2}-1}$ times larger than the ℓ_p Lewis weight, and this way of bounding the entries of $\mathbf{z}$ is one central step to avoiding the $\tilde{O}(d^{p/2})$ dependence.

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

LEMMA 5.3. *Consider Setting 5.1, and let*

$$\mathcal{B} = \left\{ i \in [n] \,:\, |\mathbf{z}(i)| > \tfrac{OPT}{\varepsilon} (w_q[\boldsymbol{B}](i))^{1/p} \right\}.$$

So that $\bar{\mathbf{z}} \in \mathbb{R}^{n_0}$ *is equal to* $\mathbf{z}$ *but with all entries in* $\mathcal{B}$ *set to zero. Then for all* $\mathbf{x} \in \mathbb{R}^{d+1}$ *with* $\|\boldsymbol{A}\mathbf{x}\|_p \le C_0\, OPT$,

$$\left| \|\boldsymbol{A}\mathbf{x} - \mathbf{z}\|_p^p - \|\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}\|_p^p - \|\mathbf{z} - \bar{\mathbf{z}}\|_p^p \right| \le C_1 \varepsilon \cdot OPT^p$$

where C_1 *is a constant that depends only on* C_0, C_z, *and* p.

Proof. For any $\mathbf{x} \in \mathbb{R}^{d+1}$, by the definition of ℓ_p sensitivity,

$$\frac{|[\boldsymbol{A}\mathbf{x}](i)|^p}{\|\boldsymbol{A}\mathbf{x}\|_p^p} \le \psi_p[\boldsymbol{A}](i)$$

From the relationship of ℓ_p sensitivities and ℓ_q sensitivities for Vandermonde matrices, i.e., Theorem 5.1, we have

$$\frac{|[\boldsymbol{A}\mathbf{x}](i)|^p}{\|\boldsymbol{A}\mathbf{x}\|_p^p} \le \psi_p[\boldsymbol{A}](i) \le \psi_q[\boldsymbol{B}](i)$$

Next, by Lemma 2.5 from [MMWY22], which says that for $q \in [0,2]$, the ℓ_q sensitivities lower bound the ℓ_q Lewis weights, we have

$$\frac{|[\boldsymbol{A}\mathbf{x}](i)|^p}{\|\boldsymbol{A}\mathbf{x}\|_p^p} \le \psi_q[\boldsymbol{B}](i) \le w_q[\boldsymbol{B}](i)$$

Thus for $i \in \mathcal{B}$ we have

$$|[\boldsymbol{A}\mathbf{x}](i)|^p \le w_q[\boldsymbol{B}](i) \cdot \|\boldsymbol{A}\mathbf{x}\|_p^p \le \varepsilon^p \|\boldsymbol{A}\mathbf{x}\|_p^p \cdot \frac{|\mathbf{z}(i)|^p}{OPT^p}$$

Since $\|\boldsymbol{A}\mathbf{x}\|_p^p \le C_0^p\, OPT^p$ by assumption, it follows that $|[\boldsymbol{A}\mathbf{x}](i)|^p \le C_0^p \varepsilon^p \cdot |\mathbf{z}(i)|^p$, and thus $|[\boldsymbol{A}\mathbf{x}](i) - \mathbf{z}(i)| \in (1 \pm C_0\varepsilon)\,|\mathbf{z}(i)|$. Using this fact and the fact that $\bar{\mathbf{z}}(i) = 0$,

$$\begin{aligned}
|[\boldsymbol{A}\mathbf{x}](i) - \mathbf{z}(i)|^p - |[\boldsymbol{A}\mathbf{x}](i) - \bar{\mathbf{z}}(i)|^p &= |[\boldsymbol{A}\mathbf{x}](i) - \mathbf{z}(i)|^p - |[\boldsymbol{A}\mathbf{x}](i)|^p \\
&\in (1 \pm C_0\varepsilon)^p\,|\mathbf{z}(i)|^p \pm C_0^p \varepsilon^p\,|\mathbf{z}(i)|^p \\
&\subseteq (1 \pm 3C_0 p\varepsilon)\,|\mathbf{z}(i)|^p \pm C_0^p \varepsilon^p\,|\mathbf{z}(i)|^p && \text{(Lemma E.4)} \\
&\subseteq (1 \pm C_0'\varepsilon)\,|\mathbf{z}(i)|^p && \text{(5.5)}
\end{aligned}$$

where the last line sets $C_0' := 3C_0 p + C_0^p$. Then, summing over all $i \in \mathcal{B}$,

$$\sum_{i\in\mathcal{B}} |[\boldsymbol{A}\mathbf{x}](i) - \mathbf{z}(i)|^p - \sum_{i\in\mathcal{B}} |[\boldsymbol{A}\mathbf{x}](i) - \bar{\mathbf{z}}(i)|^p - \sum_{i\in\mathcal{B}} |\mathbf{z}(i) - \bar{\mathbf{z}}(i)|^p \le C_0'\varepsilon \cdot \sum_{i\in\mathcal{B}} |\mathbf{z}(i)|^p$$

We have by assumption that $\sum_{i\in\mathcal{B}} |\mathbf{z}(i)|^p \le \|\mathbf{z}\|_p^p \le C_0^p\, OPT^p$. Finally, since $\bar{\mathbf{z}}(i) = \mathbf{z}(i)$ for $i \notin \mathcal{B}$, we conclude that

$$\left| \|\boldsymbol{A}\mathbf{x} - \mathbf{z}\|_p^p - \|\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}\|_p^p - \|\mathbf{z} - \bar{\mathbf{z}}\|_p^p \right| = C_0' C_z^p \varepsilon \cdot OPT^p.$$

∎

Next, we show the same intuition about $\mathbf{z}$ and $\bar{\mathbf{z}}$ holds when looking at the subsampled regression problem; that minimizing $\|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\mathbf{z}\|_p$ is roughly equivalent to minimizing $\|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\bar{\mathbf{z}}\|_p$.

LEMMA 5.4. *Consider Setting 5.1. Then with probability at least* $\frac{99}{100}$, $\|\boldsymbol{S}\mathbf{z}\|_p \le C_s OPT$ *and further for all* $\mathbf{x} \in \mathbb{R}^{d+1}$ *with* $\|\boldsymbol{A}\mathbf{x}\|_p \le C_0 OPT$, *we get*

$$\left| \|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\mathbf{z}\|_p^p - \|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\bar{\mathbf{z}}\|_p^p - \|\boldsymbol{S}(\mathbf{z} - \bar{\mathbf{z}})\|_p^p \right| \le C_2 \varepsilon \cdot OPT^p$$

where C_s *and* C_2 *are constants that depend only on* C_0, p, *and* C_z.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Proof. The proof builds off of Lemma 5.3. For any $\mathbf{x} \in \mathbb{R}^{d+1}$ and $i \in \mathcal{B}$, by multiplying both sides of Equation 5.5 by S_{ii}^p, we have that for all $i \in \mathcal{B}$,

$$|[\boldsymbol{S}\boldsymbol{A}\mathbf{x}](i) - [\boldsymbol{S}\mathbf{z}](i)|^p - |[\boldsymbol{S}\boldsymbol{A}\mathbf{x}](i) - [\boldsymbol{S}\bar{\mathbf{z}}](i)|^p \in (1 \pm C_0'\varepsilon)\,|[\boldsymbol{S}\mathbf{z}](i)|^p$$

For all $i \notin \mathcal{B}$, $|[\boldsymbol{S}\boldsymbol{A}\mathbf{x}](i) - [\boldsymbol{S}\mathbf{z}](i)|^p - |[\boldsymbol{S}\boldsymbol{A}\mathbf{x}](i) - [\boldsymbol{S}\bar{\mathbf{z}}](i)|^p = 0$. since $\bar{\mathbf{z}}(i) = \mathbf{z}(i)$ for $i \notin \mathcal{B}$. Summing over all $i \in [n_0]$, we get

$$\|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\mathbf{z}\|_p^p - \|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\bar{\mathbf{z}}\|_p^p - \|\boldsymbol{S}(\mathbf{z} - \bar{\mathbf{z}})\|_p^p \in \pm C_0'\varepsilon\|\boldsymbol{S}(\mathbf{z} - \bar{\mathbf{z}})\|_p^p$$

Next, since $\bar{\mathbf{z}}$ is a partial zeroing of $\mathbf{z}$, and since $\mathbb{E}[\|\boldsymbol{S}\mathbf{z}\|_p^p] = \|\mathbf{z}\|_p^p$ we can use Markov's inequality to bound $\|\boldsymbol{S}(\bar{\mathbf{z}} - \mathbf{z})\|_p^p \le \|\boldsymbol{S}\mathbf{z}\|_p^p \le 100\|\mathbf{z}\|_p^p \le 100C_z^p OPT^p$, with probability $\frac{99}{100}$. We conclude:

$$\|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\mathbf{z}\|_p^p - \|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\bar{\mathbf{z}}\|_p^p - \|\boldsymbol{S}(\mathbf{z} - \bar{\mathbf{z}})\|_p^p \in \pm 100C_0'C_z\varepsilon \cdot OPT$$

∎

Next we state our core technical contribution: the *Affine Embedding* guarantee. While the prior work proves this same result, they require $\tilde{O}(d^{p/2})$ samples to do so. In Section 5.3, we show that Vandermonde matrices can do this by taking $\tilde{O}(d)$ samples with probabilities proportional to the ℓ_q Lewis weights of $\boldsymbol{B}$.

THEOREM 5.2. (AFFINE EMBEDDING) *Consider Setting 5.1. Then with probability $\frac{99}{100}$, for all $\mathbf{x} \in \mathbb{R}^{d+1}$ with $\|\boldsymbol{A}\mathbf{x}\|_p \le C_0 OPT$, we have*

$$\left|\|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\bar{\mathbf{z}}\|_p^p - \|\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}\|_p^p\right| \le C_3\varepsilon \cdot OPT^p \tag{5.6}$$

where C_3 is a constant that depends only on C_0, C_z, and p.

We prove Theorem 5.2 later, in Section 5.3, and instead first show that this affine embedding suffices to prove the correctness of the overall algorithm.

THEOREM 5.3. *Consider Setting 5.1. Then, Algorithm 3 reads $O(\frac{d}{\varepsilon^{6.5p+2}}\,\mathrm{polylog}(\frac{d}{\varepsilon}))$ entries of $\mathbf{b}$ and outputs a vector $\tilde{\mathbf{x}}$ such that with probability 0.9,*

$$\|\boldsymbol{A}\tilde{\mathbf{x}} - \mathbf{b}\|_p \le (1+\varepsilon) \min_{\mathbf{x} \in \mathbb{R}^{d+1}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p$$

Proof. By Lemma 5.2, we know that step 1 from Algorithm 3 returns a vector $\mathbf{x}_c$ such that $\|\boldsymbol{A}\mathbf{x}_c - \mathbf{b}\|_p \le C_z \cdot OPT$. Recall that $\mathbf{z} := \mathbf{b} - \boldsymbol{A}\mathbf{x}_c$, so we equivalently have $\|\mathbf{z}\|_p \le C_z \cdot OPT$. Let $\bar{\mathbf{z}} \in \mathbb{R}^{n_0}$ be the partially zeroed out copy of $\mathbf{z}$ as in Setting 5.1. Combining Lemma 5.3, Lemma 5.4, and Theorem 5.2, for all $\mathbf{x}$ with $\|\boldsymbol{A}\mathbf{x}\|_p \le C_0 \cdot OPT$, we get

$$\begin{aligned}
\|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\mathbf{z}\|_p^p &\in \|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\bar{\mathbf{z}}\|_p^p + \|\boldsymbol{S}\mathbf{z} - \boldsymbol{S}\bar{\mathbf{z}}\|_p^p \pm C_2\varepsilon \cdot OPT^p && \text{(Lemma 5.4)}\\
&\subseteq \|\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}\|_p^p + \|\boldsymbol{S}\mathbf{z} - \boldsymbol{S}\bar{\mathbf{z}}\|_p^p \pm (C_2 + C_3)\varepsilon \cdot OPT^p && \text{(Theorem 5.2)}\\
&\subseteq \|\boldsymbol{A}\mathbf{x} - \mathbf{z}\|_p^p - \|\mathbf{z} - \bar{\mathbf{z}}\|_p^p + \|\boldsymbol{S}\mathbf{z} - \boldsymbol{S}\bar{\mathbf{z}}\|_p^p \pm (C_1 + C_2 + C_3)\varepsilon \cdot OPT^p && \text{(Lemma 5.3)}\\
&= \|\boldsymbol{A}\mathbf{x} - \mathbf{z}\|_p^p - \hat{C} \pm (C_1 + C_2 + C_3)\varepsilon \cdot OPT^p
\end{aligned}$$

where $\hat{C} := \|\mathbf{z} - \bar{\mathbf{z}}\|_p^p - \|\boldsymbol{S}\mathbf{z} - \boldsymbol{S}\bar{\mathbf{z}}\|_p^p$ is independent of $\mathbf{x}$. Note that since $\bar{\mathbf{z}}$ is a partial zeroing of $\mathbf{z}$, $\|\mathbf{z} - \bar{\mathbf{z}}\|_p \le \|\mathbf{z}\|_p \le C_z \cdot OPT$. Similarly, $\|\boldsymbol{S}\mathbf{z} - \boldsymbol{S}\bar{\mathbf{z}}\|_p \le \|\boldsymbol{S}\mathbf{z}\|_p \le C_s \cdot OPT$. So, we have $\hat{C} \le (C_z^p + C_s^p)\,OPT$ and thus we can equivalently write this last bound as, for any $\mathbf{x}$ with $\|\boldsymbol{A}\mathbf{x}\|_p \le C_0 \cdot OPT$,

$$\left|\|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\mathbf{z}\|_p^p - (\|\boldsymbol{A}\mathbf{x} - \mathbf{z}\|_p^p + \hat{C})\right| \le C_4\varepsilon \cdot OPT^p \tag{5.7}$$

where $C_4 := C_1 + C_2 + C_3$. We will apply Equation 5.7 twice, once to $\hat{\mathbf{x}} = \mathrm{argmin}_{\mathbf{x}} \|\boldsymbol{S}\boldsymbol{A}\mathbf{x} - \boldsymbol{S}\mathbf{z}\|_p$ and once to $\mathbf{x}^* := \mathrm{argmin}_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{z}\|_p$. To do so, we first have to verify that $\|\boldsymbol{A}\hat{\mathbf{x}}\|_p$ and $\|\boldsymbol{A}\mathbf{x}^*\|_p$ are small enough – i.e. are at most $C_0 OPT$. We first bound $\|\boldsymbol{A}\mathbf{x}^*\|_p \le \|\boldsymbol{A}\mathbf{x}^* - \bar{\mathbf{z}}\|_p + \|\bar{\mathbf{z}}\|_p \le 2\|\bar{\mathbf{z}}\|_p \le 2C_z OPT \le C_0 OPT$. Next, by

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Lemma 5.1, we have that $\boldsymbol{S}$ is an ℓ_p subspace embedding. So, we have $\|\boldsymbol{A}\hat{\mathbf{x}}\|_p \leq 2\|\boldsymbol{S}\boldsymbol{A}\hat{\mathbf{x}}\|_p$ and by Markov's inequality, with probability $\frac{99}{100}$, we have:

$$2\|\boldsymbol{S}\boldsymbol{A}\hat{\mathbf{x}}\|_p \leq 2\|\boldsymbol{S}\boldsymbol{A}\hat{\mathbf{x}} - \boldsymbol{S}\mathbf{z}\|_p + 2\|\boldsymbol{S}\mathbf{z}\|_p \leq 2\|\boldsymbol{S}\boldsymbol{A}\mathbf{x}^* - \boldsymbol{S}\mathbf{z}\|_p + 2\|\boldsymbol{S}\mathbf{z}\|_p \leq 200(\|\boldsymbol{A}\mathbf{x}^* - \mathbf{z}\|_p + \|\mathbf{z}\|_p) \leq C_0 OPT$$

We proceed to apply Equation 5.7 twice, to get

$$\begin{aligned}
\text{(Equation 5.7)} \quad \|\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{z}\|_p^p &\leq \|\boldsymbol{S}\boldsymbol{A}\hat{\mathbf{x}} - \boldsymbol{S}\mathbf{z}\|_p^p - \hat{C} + C_4\varepsilon OPT^p \\
\text{(Optimality of } \hat{\mathbf{x}}\text{)} \quad &\leq \|\boldsymbol{S}\boldsymbol{A}\mathbf{x}^* - \boldsymbol{S}\mathbf{z}\|_p^p - \hat{C} + C_4\varepsilon OPT^p \\
\text{(Equation 5.7)} \quad &\leq (\|\boldsymbol{A}\mathbf{x}^* - \mathbf{z}\|_p^p + \hat{C}) - \hat{C} + 2C_4\varepsilon OPT^p \\
&= \|\boldsymbol{A}\mathbf{x}^* - \mathbf{z}\|_p^p + 2C_4\varepsilon OPT^p
\end{aligned}$$

And so the overall predictor $\tilde{\mathbf{x}} = \mathbf{x}_c + \hat{\mathbf{x}}$ has

$$\begin{aligned}
\|\boldsymbol{A}\tilde{\mathbf{x}} - \mathbf{b}\|_p^p &= \|\boldsymbol{A}\mathbf{x}_c + \boldsymbol{A}\hat{\mathbf{x}} - \mathbf{z} - \boldsymbol{A}\mathbf{x}_c\|_p^p \\
&= \|\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{z}\|_p^p \\
&\leq \min_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{z}\|_p^p + 2C_4\varepsilon OPT^p \\
&= \min_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - (\mathbf{b} + \boldsymbol{A}\mathbf{x}_c)\|_p^p + 2C_4\varepsilon OPT^p \\
&= \min_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p^p + 2C_4\varepsilon OPT^p \\
&= (1 + 2C_4\varepsilon) \min_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p^p
\end{aligned}$$

Note that our proof ensures that Theorem 5.3 holds with a fixed constant probability. ∎

5.3 Proving the Affine Embedding (Theorem 5.2) To prove Theorem 5.2, we want a bound over all vectors $\boldsymbol{A}\mathbf{x}$ where $\|\boldsymbol{A}\mathbf{x}\|_p^p \leq C_0^p OPT^p$. Since a naïve ε-net argument would lead to a suboptimal dependence on d, we follow the first arXiv version of [MMWY22], and appeal to a more refined net analysis introduced in [BLM89]. In that work, the authors construct a "compact rounding" for all vectors in the set $\{\boldsymbol{A}\mathbf{x} : \|\boldsymbol{A}\mathbf{x}\|_p^p \leq 1\}$. In particular, they construct a series of nets $\mathcal{D}_0, \ldots, \mathcal{D}_\ell$ (with different properties for each $k \in \{0, \ldots, \ell\}$), such that every $\boldsymbol{A}\mathbf{x}$ with $\|\boldsymbol{A}\mathbf{x}\|_p^p \leq 1$ can be approximated as $\boldsymbol{A}\mathbf{x} \approx \sum_{k=0}^{\ell} \mathbf{d}_k$, where each $\mathbf{d}_k \in \mathcal{D}_k$. After scaling these vectors by a factor of $C_0 OPT$ and applying a union bound over the nets $\mathcal{D}_0, \ldots, \mathcal{D}_\ell$, [BLM89] obtains a sampling result for ℓ_p Lewis weights with an optimal d dependence of $\tilde{O}(d^{\max\{1,p/2\}})$.

To avoid this large d dependence for $p > 2$, we return to the expanded Vandermonde matrix $\boldsymbol{B} \in \mathbb{R}^{n_0 \times d_B}$. In Lemma 5.6, we show how to use the nets $\mathcal{D}_0, \ldots, \mathcal{D}_\ell$ from the ℓ_q compact rounding on $\boldsymbol{B}$ to create nets $\mathcal{E}_0, \ldots, \mathcal{E}_\ell$ for an ℓ_p compact rounding on $\boldsymbol{A}$. Each ℓ_p net $\mathcal{E}_k$ will have the same cardinality as the corresponding ℓ_q net $\mathcal{D}_k$, which makes it significantly smaller than the black-box net that would be created for Lewis weight sampling general matrices in the ℓ_p norm. Lastly, Lemma 5.6 also uses a technique from [BLM89] to transform $\mathcal{E}_0, \ldots, \mathcal{E}_\ell$, which approximate vectors of the form $\boldsymbol{A}\mathbf{x}$, into new nets $\mathcal{F}_0, \ldots, \mathcal{F}_\ell$, which have similar size and approximate vectors of the form $\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}$.

To get started, we use the following compact rounding lemma, proven in the first version of [MMWY22] uploaded to arXiv, with a complete and simplified proof included in Appendix D for completeness. Specifically, we state the result from Appendix D in the special case when $\mathbf{v} = 0$:

Lemma 5.5. (Compact Rounding, [MMWY22]) *Let $\boldsymbol{B} \in \mathbb{R}^{n_0 \times d_B}$ and $q \in [0, 2]$. Let $\mathcal{N}_{\varepsilon_c}$ be an ε_c-Net over $\|\boldsymbol{B}\mathbf{y}\|_q = 1$ with $|\mathcal{N}_\varepsilon| \leq O(d \log(\frac{1}{\varepsilon}))$. Let $\ell = \log_{1+\varepsilon_c}((2d_B)^{1/q})$. Then, there exists sets of vectors $\mathcal{D}_0, \ldots, \mathcal{D}_\ell \subseteq \mathbb{R}^{n_0}$, such that: For all $\mathbf{u} \in \mathcal{N}_{\varepsilon_c}$ we can pick $\mathbf{d}_0 \in \mathcal{D}_0, \ldots, \mathbf{d}_\ell \in \mathcal{D}_\ell$ to create a "compact rounding" $\mathbf{u}' = \sum_{k=0}^{\ell} \mathbf{d}_k$ where:*

1. $|\mathbf{u}(i) - \mathbf{u}'(i)| \leq \varepsilon_c |\mathbf{u}(i)|$ *for all* $i \in [n_0]$
2. $|\mathbf{d}_k(i)| \leq \frac{1}{\varepsilon_c}(\frac{1}{2}(\frac{w_q[\boldsymbol{B}](i)}{d_B} + \frac{1}{n_0}))^{1/q}(1 + \varepsilon_c)^{k+2}$ *for all* $i \in [n_0], k \in \{0, \ldots, \ell\}$
3. $\mathbf{d}_0, \ldots, \mathbf{d}_\ell$ *all have disjoints supports*

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Further, we have that the sets $\mathcal{D}_0, \ldots, \mathcal{D}_\ell$ are not too large:

$$\log |\mathcal{D}_k| \leq C_q \frac{d_B \log(n_0)}{\varepsilon_c^{1+q}(1+\varepsilon_c)^{qk}},$$

where C_q is a fixed constant depending only on q.

Note that we can upper bound $\frac{w_q[\boldsymbol{B}](i)}{d_B} + \frac{1}{n_0} \leq \frac{w_q[\boldsymbol{B}](i)}{d_B}\operatorname{polylog}(d)$, so we instead have $|\mathbf{d}_k(i)| \leq \frac{1}{\varepsilon}(\frac{w_q[\boldsymbol{B}](i)}{d_B})^{1/q}(1+\varepsilon)^{k+2}\operatorname{polylog}(d)$. To do this, note that by Theorem 4.4, $w_q[\boldsymbol{B}](i) \geq \frac{1}{n_0} w'(s_i) \cdot \frac{1}{\operatorname{polylog}(d)}$, where w' is the clipped Chebyshev measure for degree rd. Then, $w_q[\boldsymbol{B}](i) \geq \frac{d_B}{n_0} \cdot \frac{1}{\operatorname{polylog}(d)}$, so that $\frac{1}{n_0} \leq \frac{w_q[\boldsymbol{B}](i)}{d_B}\operatorname{polylog}(d)$, and so $\frac{w_q[\boldsymbol{B}](i)}{d_b} + \frac{1}{n_0} \leq \frac{w_q[\boldsymbol{B}](i)}{d_B}\operatorname{polylog}(d)$.

LEMMA 5.6. (VANDERMONDE COMPACT ROUNDING) *Let $\boldsymbol{A} \in \mathbb{R}^{n_0 \times d+1}$ and $p > 2$. Let $\mathcal{N}_\varepsilon$ be an ε-Net over $\|\boldsymbol{A}\mathbf{x}\|_p \leq C_0 OPT$, so that any $\mathbf{x}$ with $\|\boldsymbol{A}\mathbf{x}\|_p \leq C_0 OPT$ has some $\mathbf{u} \in \mathcal{N}_\varepsilon$ such that $\|\boldsymbol{A}\mathbf{x} - \mathbf{u}\|_p \leq \varepsilon OPT$. Then we can pick an odd integer r such that $\frac{1}{2}p \leq r \leq \frac{3}{2}p$, and let $q = \frac{p}{r} \in [\frac{2}{3}, 2]$. Let $\ell = \log_{1+\varepsilon}((2d_B)^{1/q})$. There exists sets of vectors $\mathcal{F}_0, \ldots, \mathcal{F}_\ell \subseteq \mathbb{R}^{n_0}$, such that: For any $\mathbf{u} \in \mathcal{N}_\varepsilon$ we let $\mathbf{r} := \mathbf{u} - \bar{\mathbf{z}}$ and we can pick $\mathbf{f}_0 \in \mathcal{F}_0, \ldots, \mathbf{f}_\ell \in \mathcal{F}_\ell$ to create a "compact rounding" $\mathbf{r}' = \sum_{k=0}^{\ell} \mathbf{f}_k$ where:*

1. $|\mathbf{r}(i) - \mathbf{r}'(i)| \leq \varepsilon \max\{|[\boldsymbol{A}\mathbf{x}](i)|, |\bar{\mathbf{z}}(i)|\}$ *for all* $i \in [n_0]$
2. $|\mathbf{f}_k(i)| \leq \frac{OPT}{\varepsilon^2}(\frac{w_q[\boldsymbol{B}](i)}{d_B}\operatorname{polylog}(d))^{1/p}(1+\varepsilon^r)^{\frac{k+2}{r}}$ *for all* $i \in [n_0], k \in \{0, \ldots, \ell\}$
3. $\mathbf{f}_0, \ldots, \mathbf{f}_\ell$ *all have disjoints supports*

Further, we have that the sets $\mathcal{F}_0, \ldots, \mathcal{F}_\ell$ are not too large:

$$\log |\mathcal{F}_k| \leq C_q \frac{d_B \log(n_0)}{\varepsilon^{r(1+q)}(1+\varepsilon^r)^{qk}}$$

Proof. Depending on the value of p, we will pick q differently. If $p \in (2,3)$, we let $r = 3$ and $q := \frac{p}{r} \in [\frac{2}{3}, 1]$. If $p \geq 3$, we let r be an odd integer such that $r \leq p < 2r$, and let $q := \frac{p}{r} \in [1,2]$. In both cases r is an odd integer, we have $p = qr$, we know that the ℓ_q Lewis weights of $\boldsymbol{B}$ are close to the Chebyshev measure, and Lemma 5.5 accepts this value of q. The rest of this paper will not distinguish between the $p \in (2,3)$ and the $p \geq 3$ cases.

Notably, the compact rounding requires being given $\mathcal{N}_q$, an ℓ_q net over $\{\boldsymbol{B}\mathbf{y} : \|\boldsymbol{B}\mathbf{y}\|_q \leq 1\}$. But we want to make sure that all $\boldsymbol{A}\mathbf{x} \in \mathcal{N}_\varepsilon$ have $(\boldsymbol{A}\mathbf{x})^r \in \mathcal{N}_q$. So, formally, let $\mathcal{N}_{q,0}$ be an arbitrary ℓ_q net for $\{\boldsymbol{B}\mathbf{y} : \|\boldsymbol{B}\mathbf{y}\|_q \leq 1\}$, and let $\mathcal{N}_{induced} := \{\boldsymbol{B}\mathbf{y} : ([\boldsymbol{A}\mathbf{x}](i))^r = [\boldsymbol{B}\mathbf{y}](i) \text{ for all } i\}$ be the mapping of $\mathcal{N}_\varepsilon$ to the range of $\boldsymbol{B}$. By Lemma 2.4 of [BLM89], we have that both $\mathcal{N}_{q,0}$ and $\mathcal{N}_\varepsilon$ have cardinality at most $(\frac{3}{\varepsilon})^d$. We then apply Lemma 5.5 on the net $\mathcal{N}_q := \mathcal{N}_{q,0} \bigcup \mathcal{N}_{induced}$ and with $\varepsilon_c = \varepsilon^r$.

Also, note that the vectors in Lemma 5.5 formally require $\|\boldsymbol{B}\mathbf{y}\|_q \leq 1$, while we have $\|\boldsymbol{B}\mathbf{y}\|_q = \|\boldsymbol{A}\mathbf{x}\|_p^r \leq (C_0 OPT)^r$. So, we scale up the vectors $\mathbf{d}_k$ returned by Lemma 5.5 by a factor of $(C_0 OPT)^r$, so that $|\mathbf{d}_k(i)| \leq \frac{(C_0 OPT)^r}{\varepsilon^r}(\frac{w_q[\boldsymbol{B}](i)}{d_B} + \frac{1}{n_0})^{1/q}(1+\varepsilon^r)^{k+2}$.

With this in place, now we fix any $\mathbf{u} \in \mathcal{N}_{induced}$, and let $\sum_{k=0}^{\ell} \mathbf{d}_k$ be the compact rounding of $\mathbf{u}$ as defined by Lemma 5.5. Using the fact that $qr = p$, we let $\alpha_i := \frac{OPT}{\varepsilon}(\frac{w_q[\boldsymbol{B}](i)}{d_B}\operatorname{polylog}(d))^{1/p}$ so that $|\mathbf{d}_k(i)| \leq \alpha_i^r(1+\varepsilon^r)^{k+2}$. We now intuitively round $\boldsymbol{A}\mathbf{x} \approx \sum_{k=0}^{\ell}(\mathbf{d}_k)^{1/r}$. We define $\mathbf{e}_0, \ldots, \mathbf{e}_\ell$ such that:

$$\mathbf{e}_k(i) := (\mathbf{d}_k(i))^{1/r}$$

so that $|\mathbf{e}_k(i)| \leq \alpha_i(1+\varepsilon^r)^{\frac{k+2}{r}}$. Using the fact that r is an odd integer, we have $\operatorname{sign}(\mathbf{e}_k(i)) = \operatorname{sign}(\mathbf{d}_k(i))$. Further, looking at the proof of the compact rounding in Appendix D with $\mathbf{v} = 0$, we see from Lemma D.5 that $\operatorname{sign}(\mathbf{d}_k(i)) = \operatorname{sign}(\mathbf{u}(i))$. So, we have that $\operatorname{sign}(\mathbf{e}_k(i)) = \operatorname{sign}(\boldsymbol{A}\mathbf{x}(i))$. This definition of $\mathbf{e}_k$ means that

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

$|\mathbf{e}_k(i)| \le \alpha_i(1+\varepsilon^r)^{\frac{k+2}{r}}$, and further that

$$\left|\boldsymbol{A}\mathbf{x}(i) - \sum_{k=0}^{\ell} \mathbf{e}_k(i)\right| = \left|(\mathbf{u}(i))^{1/r} - (\mathbf{d}_k(i))^{1/r}\right|$$

(The signs are equal) $= \left||\mathbf{u}(i)|^{1/r} - |\mathbf{d}_k(i)|^{1/r}\right|$

($\left|x^{1/r} - y^{1/r}\right| \le |x-y|^{1/r}$ for all $x, y \ge 0$) $\le ||\mathbf{u}(i)| - |\mathbf{d}_k(i)||^{1/r}$

$\le (\varepsilon^r |\mathbf{u}(i)|)^{1/r}$

$= \varepsilon |\boldsymbol{A}\mathbf{x}(i)|$

Also note that $\mathbf{e}_k$ has the same support as $\mathbf{d}_k$, so that all the properties of Lemma 5.5 are preserved, just in estimating a slightly different vector. We next examine rounding $\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}$ to a compact rounding. Borrowing a proof strategy from Appendix D,

(for $k \in \{0, \dots, \ell\}$) $$B_{k,\mathbf{u}} := \left\{ i \in [n_0] \,:\, \mathbf{e}_k(i) \neq 0\,,\, \varepsilon\,|\bar{\mathbf{z}}(i)| \le 2\alpha_i(1+\varepsilon^r)^{\frac{k+2}{r}} \right\}$$

(for $k \in \{1, \dots, \ell\}$) $$H_k := \left\{ i \in [n_0] \,:\, 2\alpha_i(1+\varepsilon^r)^{\frac{k+1}{r}} < \varepsilon\,|\bar{\mathbf{z}}(i)| \le 2\alpha_i(1+\varepsilon^r)^{\frac{k+2}{r}} \right\}$$

(for $k \in \{1, \dots, \ell\}$) $$G_{k,\mathbf{u}} := H_k \setminus \bigcup_{k' \ge k} \{i \in [n_0] \,:\, \mathbf{e}_{k'}(i) \neq 0\}$$

Note that $|\bar{\mathbf{z}}(i)| \le \frac{OPT}{\varepsilon}(w_q[\boldsymbol{B}](i))^{1/p} < 2\alpha_i(1+\varepsilon^r)^{\frac{\ell+2}{r}}$, so all entries of $\bar{\mathbf{z}}$ are covered by our disjoint sets. The sets $B_{0,\mathbf{u}}, \dots, B_{\ell,\mathbf{u}}, G_{1,\mathbf{u}}, \dots, G_{\ell,\mathbf{u}}$ will define the support of the final compact rounding vectors we will return, so we first show that these sets partition $[n_0]$: In the following cases, consider any k, k':

- $B_{k,\mathbf{u}} \bigcap B_{k',\mathbf{u}} = \emptyset$ since $i \in B_{k,\mathbf{u}}$ implies $\mathbf{e}_k(i) \neq 0$ implies $\mathbf{e}_{k'}(i) = 0$ implies $i \notin B_{k',\mathbf{u}}$.
- $G_{k,\mathbf{u}} \bigcap G_{k',\mathbf{u}} \subseteq H_k \bigcap H_{k'} = \emptyset$ since H_k and $H_{k'}$ have no intersection by definition.
- For $k \ge k'$, $B_{k,\mathbf{u}} \bigcap G_{k',\mathbf{u}} = \emptyset$ since $i \in B_{k,\mathbf{u}}$ means $\mathbf{e}_k(i) \neq 0$ so $i \in \bigcup_{k' \ge k}\{i \in [n_0] \,:\, \mathbf{e}_{k'}(i) \neq 0\}$ so $i \notin G_{k',\mathbf{u}}$.
- For $k < k'$, $B_{k,\mathbf{u}} \bigcap G_{k',\mathbf{u}} = \emptyset$ since $k' \ge k+1$ and $i \in H_{k'}$ means $\varepsilon\,|\bar{\mathbf{z}}(i)| > 2\alpha(1+\varepsilon^r)^{\frac{k'+1}{r}} \ge 2\alpha(1+\varepsilon^r)^{\frac{k+2}{r}}$, which contradicts $i \in B_{k,\mathbf{u}}$.

So, we can now define the vectors $\mathbf{f}_0, \dots, \mathbf{f}_\ell$ as

$$\mathbf{f}_k(i) := \begin{cases} \mathbf{e}_k(i) - \bar{\mathbf{z}}(i) & i \in B_{k,\mathbf{u}} \\ -\bar{\mathbf{z}}(i) & i \in G_{k,\mathbf{u}} \\ 0 & \textit{otherwise} \end{cases}$$

Now, we show that $\mathbf{r}' := \sum_{k=0}^{\ell} \mathbf{f}_k$ satisfies the guarantees of Lemma 5.6. Fix any $i \in [n_0]$ and let $k \in \{0, \dots, \ell\}$ be the index[10] where $\mathbf{f}_k(i) \neq 0$. Then, recalling that $\mathbf{r} = \boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}$,

(when $i \in B_{k,\mathbf{u}}$) $$|\mathbf{f}_k(i) - \mathbf{r}(i)| = |\mathbf{e}_k(i) - [\boldsymbol{A}\mathbf{x}](i)| \le \varepsilon\,|[\boldsymbol{A}\mathbf{x}](i)|$$

(when $i \in G_{k,\mathbf{u}}$) $|\mathbf{f}_k(i) - \mathbf{r}(i)| = |[\boldsymbol{A}\mathbf{x}](i)|$

(for some $k' < k$, by def of $G_{k,\mathbf{u}}$) $\le (1+\varepsilon)\,|\mathbf{e}_{k'}(i)|$

($|\mathbf{e}_{k'}(i)| \le \alpha_i(1+\varepsilon)^{\frac{k'+2}{r}}$) $\le 2\alpha_i(1+\varepsilon^r)^{\frac{k'+2}{r}}$

(def of H_k) $\le \varepsilon\,|\bar{\mathbf{z}}(i)|$

[10] Technically, we don't guarantee that all $i \in [n_0]$ are associated with some $k \in \{0, \dots, \ell\}$. But the relative error guarantee from Lemma 5.5 and definitions of $B_{k,\mathbf{u}}$ and $G_{k,\mathbf{u}}$ imply that if $\mathbf{u}(i) \neq 0$ or $\bar{\mathbf{z}}(i) \neq 0$ then such a k exists, which suffices to prove our error guarantee.

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

And so we find $|\mathbf{r}'(i) - \mathbf{r}(i)| \le \varepsilon \max\{|[\boldsymbol{A}\mathbf{x}](i)|, |\bar{\mathbf{z}}(i)|\}$. We also have that $\mathbf{f}_0, \dots, \mathbf{f}_\ell$ have disjoint supports because $B_{0,\mathbf{u}}, \dots, B_{\ell,\mathbf{u}}, G_{1,\mathbf{u}}, \dots, G_{\ell,\mathbf{u}}$ have disjoint supports.

Next, we bound the size of entries of $\mathbf{f}_k$. We have $|\mathbf{f}_k(i)| \le |\mathbf{e}(i)| + |\bar{\mathbf{z}}(i)| \le (1 + \frac{2}{\varepsilon})\alpha_i(1+\varepsilon^r)^{\frac{k+2}{r}} \le \frac{OPT}{\varepsilon^2}(\frac{w_q[\boldsymbol{B}](i)}{d_B}\operatorname{polylog}(d))^{1/p}(1+\varepsilon^r)^{\frac{k+2}{r}}$.

To bound the number of possible $\mathbf{f}_k$ vectors, note that $\mathbf{f}_k$ is a deterministic function in $B_{k,\mathbf{u}}$ and $G_{k,\mathbf{u}}$. So, let $B_k := \{B_{k,\mathbf{u}} : \mathbf{u} \in \mathcal{N}_\varepsilon\}$ be the set of all possible "B" index sets generated at layer k, and similarly let $G_k := \{B_{k,\mathbf{u}} : \mathbf{u} \in \mathcal{N}_\varepsilon\}$. Then, looking across all possible fixings of $\mathbf{u} \in \mathcal{N}_\varepsilon$, each $\mathbf{f}_k$ is deterministic in some $\mathcal{S}_1 \in B_k$ and some $\mathcal{S}_2 \in G_k$. So, the number of possible $\mathbf{f}_k$ is at most

$$|\mathcal{F}_k| = |\{\mathbf{f}_k : \mathbf{u} \in \mathcal{N}_\varepsilon\}| \le |\{(\mathcal{S}_1, \mathcal{S}_2) : \mathcal{S}_1 \in B_k,\ \mathcal{S}_2 \in G_k\}| = |B_k| \cdot |G_k|$$

Next, since $B_{k,\mathbf{u}} \subseteq \{i \in [n_0] : \mathbf{e}_k(i) \ne 0\}$, and since $\mathbf{e}_k$ are a simple bijection with $\mathbf{d}_k \in \mathcal{D}_k$, we have $|B_k| \le |\mathcal{D}_k|$. The same holds for G_k, so $|G_k| \le |\mathcal{D}_k|$. We conclude that

$$\log|\mathcal{F}_k| \le \log|B_k| + \log|G_k| \le 2\log|\mathcal{D}_k| = 2C_r\frac{d_B\log(n_0)}{\varepsilon^{r(1+q)}(1+\varepsilon^r)^{qk}}$$

□

LEMMA 5.7. *Let $p_i := \min\{1, \frac{m}{n_0}\frac{1}{\sqrt{1-s_i^2}}\}$, where $s_1, \dots, s_{n_0}$ are times samples uniformly at random from $[-1, 1]$, and where $m = O(\frac{d}{\varepsilon^{6.5p+2}}\log(d))$. Consider the diagonal sampling matrix $\boldsymbol{S} \in \mathbb{R}^{n_0 \times n_0}$ which takes $\boldsymbol{S}_{ii}^p = \frac{1}{p_i}$ with probability p_i and $\boldsymbol{S}_{ii} = 0$ otherwise. Then consider the set of all possible rounding vectors $\mathbf{r}'$ created by Lemma 5.6. With probability $\frac{98}{100}$, all such $\mathbf{r}'$ have $\|\boldsymbol{S}\mathbf{r}'\|_p^p \in \|\mathbf{r}'\| \pm \varepsilon^p OPT^p$.*

Proof. First, we simplify the probabilities p_i. We know by Lemma E.5 that $\max_i \frac{1}{\sqrt{1-s_i^2}} \le C_c\sqrt{n_0}$ with probability $\frac{99}{100}$. So,

$$\frac{m}{n_0}\frac{1}{\sqrt{1-s_i^2}} \le \frac{mC_c}{\sqrt{n_0}} \le 1$$

Where the last inequality holds so long as $m \le O(\sqrt{n_0}) = \tilde{O}(d^{2.5}p^{O(p)}\frac{1}{\varepsilon^{O(p^2)}})$, which is satisfied by our choice of m. This means that $p_i = \min\{1, \frac{m}{n_0}\frac{1}{\sqrt{1-s_i^2}}\}$ can be simplified to just $p_i = \frac{m}{n_0}\frac{1}{\sqrt{1-s_i^2}}$.

Now, we move onto proving the correctness of $\|\boldsymbol{S}\mathbf{r}'\|_p^p$. Fix any compact rounding $\mathbf{r}' = \sum_{k=0}^{\ell}\mathbf{f}_k$ created by Lemma 5.6. Then, since $\mathbf{f}_0, \dots, \mathbf{f}_\ell$ have disjoint support,

$$\|\boldsymbol{S}\mathbf{r}\|_p^p = \sum_{k=0}^{\ell}\|\boldsymbol{S}\mathbf{f}_k\|_p^p$$

So it suffices to just prove that $\|\boldsymbol{S}\mathbf{f}_k\|_p^p \in \|\mathbf{f}_k\|_p^p \pm \frac{\varepsilon^p}{\ell+1}OPT^p$ for all $\mathbf{f}_k \in \mathcal{F}_k$ for all $k \in \{0, \dots, \ell\}$. The rest of this proof shows this concentration across all $\mathbf{f}_k$ vectors.

Fix any $\mathbf{f}_k \in \mathcal{F}_k$ for any $k \in \{0, \dots, \ell\}$. Then, we have:

(Lemma 5.6) $$|\mathbf{f}_k(i)|^p \le \frac{w_q[\boldsymbol{B}](i)}{d_B}\cdot\frac{(1+\varepsilon^r)^{q(k+2)}}{\varepsilon^{2p}}OPT^p\operatorname{polylog}(d)$$

$$\frac{1}{p_i}|\mathbf{f}_k(i)|^p \le \frac{1}{p_i}\frac{w_q[\boldsymbol{B}](i)}{d_B}\cdot\frac{(1+\varepsilon^r)^{q(k+2)}}{\varepsilon^{2p}}OPT^p\operatorname{polylog}(d)$$

($p_i = \frac{m}{n_0\sqrt{1-s_i^2}}$) $$= \frac{n_0\sqrt{1-s_i^2}}{md_B}\cdot w_q[\boldsymbol{B}](i)\cdot\frac{(1+\varepsilon^r)^{q(k+2)}}{\varepsilon^{2p}}OPT^p\operatorname{polylog}(d)$$

(Theorem 4.4) $$= \frac{n_0\sqrt{1-s_i^2}}{md_B}\cdot\left(\frac{d_B}{n_0}\frac{1}{\sqrt{1-s_i^2}}\operatorname{polylog}(d)\right)\cdot\frac{(1+\varepsilon^r)^{q(k+2)}}{\varepsilon^{2p}}OPT^p\operatorname{polylog}(d)$$

$$= \frac{(1+\varepsilon^r)^{q(k+2)}}{m\varepsilon^{2p}}OPT^p\operatorname{polylog}(d)$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Next, we will let $X_i := \boldsymbol{S}_{ii}^p |\mathbf{f}_k(i)|^p - |\mathbf{f}_k(i)|^p$, which are mean-zero random variables such that $\sum_{i=1}^{n_0} X_i = \|\boldsymbol{S}\mathbf{f}_k\|_p^p - \|\mathbf{f}_k\|_p^p$. Letting $B(n,p)$ be the binomial distribution, we then bound

$$\mathbb{E}[X_i^2] = \operatorname{Var}[\boldsymbol{S}_{ii}^p |\mathbf{f}_k(i)|^p] = \tfrac{1}{p_i^2} |\mathbf{f}_k(i)|^{2p} \operatorname{Var}[B(1,p_i)] \le \tfrac{1}{p_i} |\mathbf{f}_k(i)|^{2p}$$
$$\sum_{i=1}^{n_0} \mathbb{E}[X_i^2] \le \sum_{i=1}^{n_0} \frac{1}{p_i} |\mathbf{f}_k(i)|^2 \le \|\mathbf{f}_k\|_p^p \cdot \max_{i\in[n_0]} \tfrac{1}{p_i} |\mathbf{f}_k(i)|^p$$

And so, by Bernstein's Inequality (Imported Theorem A.1) and since $|X_i| \le \max_i \frac{1}{p_i} |\mathbf{f}_k(i)|^p$, we get the concentration

$$\begin{aligned}\Pr[\big|\|\boldsymbol{S}\mathbf{f}_k\|_p^p - \|\mathbf{f}_k\|_p^p\big| \le \gamma\, OPT^p] &= \Pr[|\textstyle\sum_{i=1}^{n_0} X_i| \le \gamma\, OPT^p] \\ &\le 2\exp\left(-\frac{\frac{1}{2}\gamma^2\, OPT^{2p}}{(\|\mathbf{f}_k\|_p^p + \frac{\gamma}{3}\, OPT^p)\cdot \max_i \frac{1}{p_i} |\mathbf{f}_k(i)|^p}\right)\end{aligned}$$

Since $\gamma = \frac{\varepsilon^p}{\ell+1} \le 1$ and $\|\mathbf{f}_k\|_p^p \le \|\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}\|_p^p \le (C_0 + C_z)^p\, OPT^p$, and letting $C_B = 2((C_0+C_z)^p + 1)$:

$$\begin{aligned}&\le 2\exp\left(-\frac{\gamma^2\, OPT^{2p}}{C_B\, OPT^p \cdot \max_i \frac{1}{p_i} |\mathbf{f}_k(i)|^p}\right) \\ &\le 2\exp\left(-\frac{\gamma^2\, OPT^p}{C_b \cdot \frac{(1+\varepsilon^r)^{q(k+2)}}{m\varepsilon^{2p}}\, OPT^p \operatorname{polylog}(d)}\right) \\ &= 2\exp\left(-m\frac{\gamma^2\varepsilon^{2p}}{(1+\varepsilon^r)^{q(k+2)}\operatorname{polylog}(d)}\right) \\ &\le \delta\end{aligned}$$

This is less than δ for $m = \frac{(1+\varepsilon^r)^{q(k+2)}}{\varepsilon^{2p}\gamma^2} \operatorname{polylog}(d) \log(\frac{2}{\delta})$. Union bounding over all $\mathbf{f}_k \in \mathcal{F}_k$, we get

$$\begin{aligned}m &= \frac{(1+\varepsilon^r)^{q(k+2)}}{\varepsilon^{2p}\gamma^2} \cdot C_r \frac{d_B \log(n_0)}{\varepsilon^{r(1+q)}(1+\varepsilon^r)^{qk}} \cdot \operatorname{polylog}(d)\log(\tfrac{2}{\delta}) \\ &= d_B \frac{(1+\varepsilon^r)^{2q}}{\varepsilon^{3p+r}\gamma^2} \cdot \log(n_0)\operatorname{polylog}(d)\log(\tfrac{2}{\delta})\end{aligned}$$

Note $(1+\varepsilon^r)^{2q} \le 2^{2q} \le 2^4$. Lastly, we union bound over all $k \in [\ell]$, where $\ell = O(\frac{\log(d)}{\varepsilon})$, so that $\gamma = \frac{\varepsilon^p}{\ell+1} = O(\frac{\varepsilon^{p+1}}{\log(d)})$, and also recall that $n_0 = O(\frac{1}{\varepsilon^{O(p^2)}} d^5 p^{O(p^2)} \log(\frac{d}{\varepsilon}))$ so that $\log(n_0) = O(p^2 \log(\frac{pd}{\varepsilon}))$, and that $r \le \frac{3}{2}p$, so we conclude that

$$m = \frac{d_B}{\varepsilon^{6.5p+2}} \cdot \operatorname{polylog}(\tfrac{pd}{\varepsilon\delta})$$

samples suffice to achieve the embeddings for all $\mathbf{f}_k$ and therefore for all $\mathbf{r}'$. ∎

LEMMA 5.8. *Let $\mathcal{N}_\varepsilon$ be an ε-Net on $\{\boldsymbol{A}\mathbf{x} : \|\boldsymbol{A}\mathbf{x}\|_p \le C_0 OPT\}$, so that for any $\boldsymbol{A}\mathbf{x}$ in this set there exists some $\mathbf{u} \in \mathcal{N}_\varepsilon$ such that $\|\boldsymbol{A}\mathbf{x} - \mathbf{u}\|_p \le \varepsilon OPT$. Consider the set of possible residual vectors $\mathbf{r} = \mathbf{u} - \bar{\mathbf{z}}$ for all $\mathbf{u} \in \mathcal{N}_\varepsilon$, and the corresponding roundings $\mathbf{r}'$ created by Lemma 5.6. Suppose the sampling matrix $\boldsymbol{S}$ ensures that $\|\boldsymbol{S}\mathbf{r}'\|_p^p \in \|\mathbf{r}'\|_p^p \pm \varepsilon OPT$. Then, $\|\boldsymbol{S}\mathbf{r}\|_p^p \in \|\mathbf{r}\|_p^p \pm C_{\mathcal{N}}\varepsilon^p \cdot OPT^p$, where $C_{\mathcal{N}}$ is a constant that depends only on C_0, C_z, and p.*

Proof. We start with a triangle inequality to show three individual terms we need to bound:

$$|\|\boldsymbol{S}\mathbf{r}\|_p - \|\mathbf{r}\|_p| \le |\|\boldsymbol{S}\mathbf{r}'\|_p - \|\mathbf{r}'\|_p| + \|\mathbf{r} - \mathbf{r}'\|_p + \|\boldsymbol{S}\mathbf{r} - \boldsymbol{S}\mathbf{r}'\|_p$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

For two numbers $b \geq a \geq 0$, we have $(b-a)^p \leq (b-a)b^{p-1} \leq b^p - ab^{p-1} \leq b^p - a^p$. So, our given assumption on $\|\boldsymbol{S}\mathbf{r}'\|_p^p$ implies that $|\|\boldsymbol{S}\mathbf{r}'\|_p - \|\mathbf{r}'\|_p|^p \leq \left|\|\boldsymbol{S}\mathbf{r}'\|_p^p - \|\mathbf{r}'\|_p^p\right| \leq \varepsilon^p OPT^p$. That is, the first term above is bounded by εOPT. The second term relies on the first property of Lemma 5.5, which bounds $|\mathbf{r}(i) - \mathbf{r}'(i)| \leq \varepsilon \max\{|\mathbf{u}(i)|, |\bar{\mathbf{z}}(i)|\}$. From there, we get

$$
\begin{aligned}
|\mathbf{r}(i) - \mathbf{r}'(i)| &\leq \varepsilon \max\{|\mathbf{u}(i)|, |\bar{\mathbf{z}}(i)|\} \\
|\mathbf{r}(i) - \mathbf{r}'(i)|^p &\leq \varepsilon^p \max\{|\mathbf{u}(i)|^p, |\bar{\mathbf{z}}(i)|^p\} \\
&\leq \varepsilon^p (|\mathbf{u}(i)|^p + |\bar{\mathbf{z}}(i)|^p) \\
\|\mathbf{r} - \mathbf{r}'\|_p^p &\leq \varepsilon^p (\|\mathbf{u}\|_p^p + \|\bar{\mathbf{z}}\|_p^p) \\
&\leq \varepsilon^p (C_0^p OPT^p + C_z^p OPT^p) \\
\|\mathbf{r} - \mathbf{r}'\|_p &\leq (C_0^p + C_z^p)^{1/p} \varepsilon OPT
\end{aligned} \tag{5.8}
$$

We lastly have to bound $\|\boldsymbol{S}\mathbf{r} - \boldsymbol{S}\mathbf{r}'\|_p$. Recall that $\boldsymbol{S}$ is a diagonal matrix. This lets us expand

$$
\begin{aligned}
\|\boldsymbol{S}(\mathbf{r} - \mathbf{r}')\|_p^p &= \sum_{i=1}^n \boldsymbol{S}_{ii} |\mathbf{r}(i) - \mathbf{r}'(i)|^p \\
&\leq \varepsilon^p \sum_{i=1}^n \boldsymbol{S}_{ii} (|\mathbf{u}(i)|^p + |\bar{\mathbf{z}}(i)|^p) && \text{(By Equation 5.8)} \\
&= \varepsilon^p (\|\boldsymbol{S}\mathbf{u}\|_p^p + \|\boldsymbol{S}\bar{\mathbf{z}}\|_p^p) \\
&\leq \varepsilon^p (2^p \|\mathbf{u}\|_p^p + 100 \|\bar{\mathbf{z}}\|_p^p) && \text{(Subspace Embedding on } \mathbf{u} \text{ and Markov's Inequality on } \bar{\mathbf{z}}) \\
&\leq \varepsilon^p (2^p C_0^p OPT^p + 100 C_z^p OPT^p) \\
\|\boldsymbol{S}(\mathbf{r} - \mathbf{r}')\|_p &\leq (2^p C_0^p + 100 C_z^p)^{1/p} \varepsilon OPT
\end{aligned}
$$

Which means we can overall bound

$$
\begin{aligned}
|\|\boldsymbol{S}\mathbf{r}\|_p - \|\mathbf{r}\|_p| &\leq |\|\boldsymbol{S}\mathbf{r}'\|_p - \|\mathbf{r}'\|_p| + \|\mathbf{r} - \mathbf{r}'\|_p + \|\boldsymbol{S}\mathbf{r} - \boldsymbol{S}\mathbf{r}'\|_p \\
&\leq (1 + 2(C_0^p + C_z^p)^{1/p} + 2(2^p C_0^p + 100 C_z^p)^{1/p)}) \varepsilon \cdot OPT
\end{aligned}
$$

□

LEMMA 5.9. *Let $\mathcal{N}_\varepsilon$ be an ε-Net on $\{\boldsymbol{A}\mathbf{x} : \|\boldsymbol{A}\mathbf{x}\|_p \leq C_0 OPT\}$, so that for any $\boldsymbol{A}\mathbf{x}$ in this set there exists some $\mathbf{u} \in \mathcal{N}_\varepsilon$ such that $\|\boldsymbol{A}\mathbf{x} - \mathbf{u}\|_p \leq \varepsilon OPT$. Consider the set of possible residual vectors $\mathbf{r} = \mathbf{u} - \bar{\mathbf{z}}$ for all $\mathbf{u} \in \mathcal{N}_\varepsilon$. Suppose the sampling matrix $\boldsymbol{S}$ ensures that $\|\boldsymbol{S}\mathbf{r}\|_p^p \in \|\mathbf{r}\|_p^p \pm C_{\mathcal{N}} \varepsilon \cdot OPT$. Then, for all $\mathbf{x}$ with $\|\boldsymbol{A}\mathbf{x}\|_p \leq C_0 OPT$, $\|\boldsymbol{S}(\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}})\|_p^p \in \|\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}\|_p^p \pm C_3 \varepsilon \cdot OPT^p$, where C_3 is a constant that depends only on $C_0, C_z, C_{\mathcal{N}}$, and p.*

Proof. Fix any $\mathbf{x}$ with $\|\boldsymbol{A}\mathbf{x}\|_p \leq C_0 OPT$. Let $\mathbf{u} \in \mathcal{N}_\varepsilon$ such that $\|\boldsymbol{A}\mathbf{x} - \mathbf{y}\|_p \leq \varepsilon OPT$. Then, by triangle inequality

$$
\begin{aligned}
|\|\boldsymbol{S}(\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}})\|_p - \|\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}\|_p| &\leq |\|\boldsymbol{S}(\mathbf{u} - \bar{\mathbf{z}})\|_p - \|\mathbf{u} - \bar{\mathbf{z}}\|_p| + \|\boldsymbol{S}(\boldsymbol{A}\mathbf{x} - \mathbf{u})\|_p + \|\boldsymbol{A}\mathbf{x} - \mathbf{u}\|_p \\
&\leq C_{\mathcal{N}} \varepsilon OPT + 3\|\boldsymbol{A}\mathbf{x} - \mathbf{u}\|_p && (\boldsymbol{S} \text{ is a subspace embedding}) \\
&\leq (C_{\mathcal{N}} + 3)\varepsilon \cdot OPT
\end{aligned}
$$

Note that for $a, b \in [0, \frac{1}{2}]$ and $p \geq 2$, we have $|a^p - b^p| \leq |a - b|$. Therefore, for any $c, d > 0$, by setting $a = \frac{c}{2\max\{c,d\}}$ and $b = \frac{d}{2\max\{c,d\}}$ and simplifying, we get $|c^p - d^p| \leq (2\max\{c,d\})^{p-1} |c - d|$. In our setting, we note that $\|\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}\|_p \leq \|\boldsymbol{A}\mathbf{x}\|_p + \|\bar{\mathbf{z}}\|_p \leq (C_0 + C_z) OPT$. Further, $\|\boldsymbol{S}(\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}})\|_p \leq \|\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}\|_p + \varepsilon(3 + C_{\mathcal{N}}) OPT \leq (C_0 + 3 + C_{\mathcal{N}}) OPT$. So, letting $C_s := (C_0 + C_z + 3 + C_{\mathcal{N}})$, we have $\max\{\|\boldsymbol{S}(\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}})\|_p, \|\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}}\|_p\} \leq C_s OPT$, and so

$$
\left|\|\boldsymbol{S}(\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}})\|_p^p - \|\boldsymbol{A}\mathbf{z} - \bar{\mathbf{z}}\|_p^p\right| \leq (2C_s OPT)^{p-1} |\|\boldsymbol{S}(\boldsymbol{A}\mathbf{x} - \bar{\mathbf{z}})\|_p - \|\boldsymbol{A}\mathbf{z} - \bar{\mathbf{z}}\|_p| \leq (2C_s)^{p-1} C' \varepsilon \cdot OPT^p
$$

□

This concludes the proof of Theorem 5.2.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

5.4 Lower Bounds for L_p Regression We now show that $(1+\varepsilon)$-approximation for L_p regression requires reading at least $\Omega\left(\frac{1}{\varepsilon^{p-1}}\right)$ entries of the function f. Later, in Section 6, we show that even 2-approximation for L_∞ regression requires reading $\Omega(n)$ entries of f.

THEOREM 5.4. *Fix $p > 1$. Any algorithm that can output a $(1+\varepsilon)$ approximation to L_p polynomial regression with probability at least $\frac{2}{3}$ must use $n = \Omega(\frac{1}{\varepsilon^{p-1}})$ queries.*

Proof. Suppose an algorithm uses $n \le \frac{1}{4\varepsilon^p}$ queries. Then there must exist an interval $\mathcal{I} \subset [-1,1]$ of width $\frac{1}{4n}$ such that none of the algorithm's queries land within $\mathcal{I}$ with probability $\frac{2}{3}$. We then define two functions:

$$f_+(t) := \begin{cases} +\frac{2^{1/p}}{\varepsilon} & t \in \mathcal{I} \\ 0 & t \notin \mathcal{I} \end{cases} \qquad f_-(t) := \begin{cases} -\frac{2^{1/p}}{\varepsilon} & t \in \mathcal{I} \\ 0 & t \notin \mathcal{I} \end{cases}$$

Both f_+ and f_- have $\|f_+\|_p^p = \|f_-\|_p^p = \frac{1}{4n} \cdot \frac{2}{\varepsilon^p}$. Let $C := 2^{-1/p} - \frac{1}{2} \in (0, \frac{1}{2})$. Then both functions have $\min_{\deg(q)\le d} \|q - f\|_p^p \le (1 - C\varepsilon)\|f\|_p^p$, since the polynomials $q_+(t) := 1$ and $q_-(t) := -1$ achieve this L_p norm:

$$\begin{aligned}
\|q_+ - f_+\|_p^p &= \tfrac{1}{4n}(\tfrac{2^{1/p}}{\varepsilon} - 1)^p + (1 - \tfrac{1}{4n})(0-1)^p \\
&\le \tfrac{1}{4n}\left((\tfrac{2^{1/p}}{\varepsilon} - 1)^p + (4n - 0)\right) \\
&= \tfrac{1}{4n}\left(\tfrac{2}{\varepsilon^p}(1 - \tfrac{\varepsilon}{2^{1/p}})^p + 4n\right) \\
&\le \tfrac{1}{4n}\left(\tfrac{2}{\varepsilon^p}(1 - \tfrac{\varepsilon}{2^{1/p}}) + \tfrac{1}{\varepsilon^p}\right) \\
&= \tfrac{1}{4n} \cdot \tfrac{2}{\varepsilon^p}\left(1 - (2^{-1/p} - \tfrac{1}{2})\,\varepsilon\right) \\
&= (1 - C\varepsilon)\|f_+\|_p^p
\end{aligned}$$

Or equivalently, $\|f_+\|_p^p \ge \frac{1}{1-C\varepsilon} \min_{\deg(q)\le d} \|q - f\|_p^p > (1 + C\varepsilon) \min_{\deg(q)\le d} \|q - f\|_p^p$. Now suppose some polynomial $\hat{q}$ has $\|\hat{q} - f_+\|_p^p \le (1-\gamma)\|f_+\|_p^p$. Since $\|f_+ - f_-\|_p = 2\|f_+\|_p$, we have

$$\begin{aligned}
\|\hat{q} - f_-\|_p &\ge \|f_+ - f_-\|_p - \|\hat{q} - f_+\|_p \\
&= 2\|f_+\|_p - (1-\gamma)^{1/p}\|f_+\|_p \\
&= (2 - (1-\gamma)^{1/p})\|f_-\|_p \\
&\ge (1 + \tfrac{\gamma}{p}\varepsilon)\|f_-\|_p \\
\|\hat{q} - f_-\|_p^p &\ge (1+\gamma)\|f_-\|_p^p
\end{aligned}$$

That is, if $\hat{q}$ is a slightly good approximation to f_+, then $\hat{q}$ is a slightly bad approximation to f_-. By symmetry, the inverse claim also holds.

To complete the argument, suppose nature picks f_+ or f_- uniformly at random. Then with probability $\frac{2}{3}$ the algorithm returns some polynomial $\hat{q}$ without knowing which function nature chose. If $\|\hat{q} - f_+\|_p^p \le \|f_+\|_p^p$ then $\|\hat{q} - f_-\|_p^p \ge \|f_-\|_p^p$, and otherwise $\|\hat{q} - f_+\|_p^p > \|f_+\|_p^p$. So, with probability $\frac{2}{3} \cdot \frac{1}{2} = \frac{1}{3}$ the resulting polynomial has error

$$\|\hat{q} - f\|_p^p \ge \|f\|_p^p > (1 + C\varepsilon) \min_{\deg(q)\le d} \|q - f\|_p^p$$

By adjusting the value of ε, we complete the proof. ∎

6 Near-Optimal L_∞ Regression

We now demonstrate how to extend these guarantees from L_p polynomial regression into L_∞ polynomial regression. We remark that the sample complexity and approximation factor guarantees in this section were already shown in [KKP17], but with a different algorithm.

For a finite dimensional regression problem with m rows, we could achieve a $(1+\varepsilon)$-approximation to ℓ_∞ regression by approximately solving ℓ_p regression with $p = \frac{\log m}{\varepsilon}$ [MMM+22]. However, since polynomials lie within an infinite dimensional space, we cannot naïvely apply this argument. In fact, it can be shown that even with arbitrarily many observations it is impossible to solve polynomial L_∞ regression to better than a 2-factor approximation:

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

THEOREM 6.1. *There does not exist an algorithm that can output a 2-approximation to L_∞ polynomial regression with probability at least $\frac{2}{3}$.*

Proof. Consider an algorithm that observes at most a finite number, say $n < \infty$, of queries from f. Then there exists some interval $\mathcal{I} \subset [-1, 1]$ of nonzero width such that none of the algorithm's queries land within $\mathcal{I}$ with probability $\frac{2}{3}$. We then define two functions:

$$f_+(t) := \begin{cases} +1 & t \in \mathcal{I} \\ 0 & t \notin \mathcal{I} \end{cases} \qquad f_-(t) := \begin{cases} -1 & t \in \mathcal{I} \\ 0 & t \notin \mathcal{I} \end{cases}$$

Both f_+ and f_- have $\|f_+\|_\infty = \|f_-\| = 1$, and both have $\min_{\deg(q) \le d} \|q - f\|_\infty \le \frac{1}{2}$, since the polynomials $q_+(t) := \frac{1}{2}$ and $q_-(t) := -\frac{1}{2}$ achieve uniform error $\frac{1}{2}$.

To complete the argument, suppose nature picks f_+ or f_- uniformly at random. Then with probability $\frac{2}{3}$ the algorithm returns some polynomial $\hat{q}$ without knowing which function nature chose. If $\hat{q}(t) \ge 0$ anywhere on $\mathcal{I}$ then $\|q - f_-\|_\infty \ge 1$, and if $\hat{q}(t) \le 0$ anywhere on $\mathcal{I}$ then $\|q - f_+\|_\infty \ge 1$. So, with probability $\frac{2}{3} \cdot \frac{1}{2} = \frac{1}{3}$ the resulting vector has error $\|\hat{q} - f\| \ge 1 \ge 2 \min_{q:\deg(q) \le d} \|q - f\|_\infty$. ∎

In light of the lower bound in Theorem 6.1, we aim to provide a constant-factor approximation for L_∞ polynomial regression rather than $(1 + \varepsilon)$-approximation. This requires a slightly different algorithm than Algorithm 1, shown below in Algorithm 5. The only changes are that the rescaling matrix now has p in the numerator, and that $\mathbf{x}$ is computed by ℓ_∞ matrix regression.

Algorithm 5 Chebyshev sampling for L_∞ polynomial regression

Input: Access to signal f, parameter $p \ge 1$, degree d, number of samples n
Output: Degree d polynomial $p(t)$
1: Sample $t_1, \dots, t_n \in [-1, 1]$ i.i.d. from the pdf $\frac{1}{\pi\sqrt{1-t^2}}$
2: Observe signal samples $b_i := f(t_i)$ for all $i \in [n]$
3: Build $\boldsymbol{A} \in \mathbb{R}^{n \times (d+1)}$ and diagonal $\boldsymbol{R} \in \mathbb{R}^{n \times n}$ with $[\boldsymbol{A}]_{i,j} = t_i^{j-1}$ and $[\boldsymbol{R}]_{ii} = \left(\frac{dp}{n}\sqrt{1 - t_i^2}\right)^{1/p}$
4: Compute $\mathbf{x} = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^{d+1}} \|\boldsymbol{RA}\mathbf{x} - \boldsymbol{R}\mathbf{b}\|_\infty$
5: Return $p(t) = \sum_{i=0}^{d} x_i t^i$

THEOREM 6.2. *Let $B(n, r)$ denote the binomial distribution. Let $n_0 = O(d^5 \operatorname{polylog} d)$ and let $p = O(\log d)$. Suppose an algorithm samples $n \sim B(n_0, 1/\tilde{O}(d^4))$ and runs Algorithm 5. Then, with probability $\frac{2}{3}$, the resulting polynomial $\hat{q}$ satisfies*

$$\|\hat{q} - f\|_\infty \le O(1) \min_{q:\deg(q) \le d} \|q - f\|_\infty$$

We prove this by mirroring a known proof technique found in Appendix A of [PPP21], which says that having a subspace embedding suffices to constant-factor approximation guarantees in any normed space. So, to apply this proof technique, we first have to have a subspace embedding in the L_∞ norm:

LEMMA 6.1. *Suppose an algorithm samples $n \sim B(d^4, O(\frac{1}{d^3}))$ and runs Algorithm 5. Then, the matrix $\boldsymbol{RA}$ on line 4 of the algorithm is a subspace embedding for $\mathcal{P}$: $\frac{1}{C}\|\mathcal{P}\mathbf{x}\|_\infty \le \|\boldsymbol{RA}\mathbf{x}\|_\infty \le C\|\mathcal{P}\mathbf{x}\|_\infty$ for all $\mathbf{x} \in \mathbb{R}^{d+1}$.*

This is the conclusion of two shorter lemmas

LEMMA 6.2. *Let $p > 2$ be an integer. Suppose an algorithm samples $n \sim B(d^4, O(\frac{1}{d^3}))$ and runs Algorithm 5. Then, the matrix $\boldsymbol{RA}$ on line 4 of the algorithm is a subspace embedding for $\mathcal{P}$: $\frac{1}{C}\|\mathcal{P}\mathbf{x}\|_p^p \le \|\boldsymbol{RA}\mathbf{x}\|_p^p \le C\|\mathcal{P}\mathbf{x}\|_p^p$ for all $\mathbf{x} \in \mathbb{R}^{d+1}$.*

Proof. We start by using the same trick as Theorem 5.1 in Section 5.2 to build a subspace embedding for large p. Let $\mathcal{Q} : \mathbb{R}^{dp+1} \to L_1([-1, 1])$ be the extended polynomial operator, so that $[\mathcal{Q}\mathbf{v}](t) = \sum_{i=0}^{dp} x_i t^i$.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Notice that $(\mathcal{P}\mathbf{x})^p$ is just some polynomial raised to integer power p. So, for any $\mathbf{x} \in \mathbb{R}^{d+1}$, there exists a $\mathbf{v} \in \mathbb{R}^{dp+1}$ such that $(\mathcal{P}\mathbf{x})^p = \mathcal{Q}\mathbf{v}$. Then, we can write

$$\|\mathcal{P}\mathbf{x}\|_p^p = \int_{-1}^{1} |[\mathcal{P}\mathbf{x}](t)|^p dt = \int_{-1}^{1} |[\mathcal{Q}\mathbf{v}](t)| dt = \|\mathcal{Q}\mathbf{v}\|_1$$

We then apply Corollary 4.2 to the L_1 norm for polynomials of degree dp. This tells us that diagonal $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ with $[\boldsymbol{S}]_{ii} = \frac{dp}{n}\sqrt{1-t_i^2}$ and Vandermonde $\boldsymbol{B} \in \mathbb{R}^{n\times(dr+1)}$ with $[\boldsymbol{B}]_{ij} = t_i^j$ enjoy

$$\frac{1}{C}\|\mathcal{Q}\mathbf{v}\|_1 \le \|\boldsymbol{S}\boldsymbol{B}\mathbf{v}\|_1 \le C\|\mathcal{Q}\mathbf{v}\|_1 \qquad \text{for all } \mathbf{v} \in \mathbb{R}^{dp+1}$$

Since $\boldsymbol{A}$ and $\boldsymbol{B}$ are just Vandermonde matrices of degree d and dp respectively, we can use the same observation to equate $\|\boldsymbol{S}\boldsymbol{B}\mathbf{v}\|_1 = \|\boldsymbol{R}\boldsymbol{A}\mathbf{x}\|_p^p$:

$$\|\boldsymbol{S}\boldsymbol{B}\mathbf{v}\|_1 = \sum_{i=1}^{n} \boldsymbol{S}_{ii}\, |[\mathcal{Q}\mathbf{v}](t_i)| = \sum_{i=1}^{n} (\boldsymbol{R}_{ii}\, |[\mathcal{P}\mathbf{x}](t_i)|)^p = \|\boldsymbol{R}\boldsymbol{A}\mathbf{x}\|_p^p$$

Where we use the fact that $[\boldsymbol{S}]_{ii} = [\boldsymbol{R}]_{ii}^p = \frac{dp}{n}\sqrt{1-t_i^2}$. So, the subspace embedding guarantee is equivalent to

$$\frac{1}{C}\|\mathcal{P}\mathbf{x}\|_p^p \le \|\boldsymbol{R}\boldsymbol{A}\mathbf{x}\|_p^p \le C\|\mathcal{P}\mathbf{x}\|_p^p \qquad \text{for all } \mathbf{x} \in \mathbb{R}^{d+1}$$

This complete the process of making a subspace embedding for large p. ∎

Next, we take $p = O(\log d)$ and show this creates an L_∞ subspace embedding.

LEMMA 6.3. *Let $p = O(\log d)$ be an integer. Suppose an algorithm samples $n \sim B(d^5, \tilde{O}(\frac{1}{d^4}))$ and runs Algorithm 5. Then, the matrix $\boldsymbol{R}\boldsymbol{A}$ on line 4 of the algorithm is a subspace embedding for $\mathcal{P}$: $\frac{1}{C}\|\mathcal{P}\mathbf{x}\|_\infty \le \|\boldsymbol{R}\boldsymbol{A}\mathbf{x}\|_\infty \le C\|\mathcal{P}\mathbf{x}\|_\infty$ for all $\mathbf{x} \in \mathbb{R}^{d+1}$.*

Proof. We achieve this by showing $\|\mathcal{P}\mathbf{x}\|_p \approx_{O(1)} \|\mathcal{P}\mathbf{x}\|_\infty$ and $\|\boldsymbol{R}\boldsymbol{A}\mathbf{x}\|_p \approx_{O(1)} \|\boldsymbol{R}\boldsymbol{A}\mathbf{x}\|_\infty$ for all $\mathbf{x}$.

This is simple to show in the finite dimensional case. By standard finite dimensional ℓ_p norm inequalities, $\|\boldsymbol{R}\boldsymbol{A}\mathbf{x}\|_\infty \le \|\boldsymbol{R}\boldsymbol{A}\mathbf{x}\|_p \le n^{\frac{1}{p}}\|\boldsymbol{R}\boldsymbol{A}\mathbf{x}\|_\infty$. Since $n = \tilde{O}(d)$, having $p = O(\log d)$ suffices for $n^{\frac{1}{p}}$ to be $O(1)$.

The infinite dimension case is more involved. We need to show that for any polynomial $h(t)$ of degree d, we have $\|h\|_\infty \approx_c \|h\|_p$. One direction is simple to show:

$$\|h\|_p^p = \int_{-1}^{1} |h(t)|^p dt \le 2\|h\|_\infty$$

The other direction follows from the Markov Brothers' Inequality, using an argument similar to Lemma 4.2. Without loss of generality assume that $\|h\|_\infty = 1$, and that $h(t_0) = 1$ for some $t_0 < 0$. Then, by Markov Brothers', we have $|h(t_0 + x)| \ge 1 - d^2 x$ for any $0 < x < \frac{1}{d^2}$. In particular, we have $|h(t)| > 1 - \frac{1}{d}$ for $t \in [t_0, t_0 + \frac{1}{d^3}]$. Then,

$$\begin{aligned}
\|h\|_p &= \left(\int_{-1}^{1} |h(t)|^p dt\right)^{1/p} \\
&\ge \left(\frac{1}{d^3}(1-\tfrac{1}{d})^p\right)^{1/p} \\
&= \frac{1}{d^{3/p}}(1-\tfrac{1}{d}) \\
&= \Omega(1)
\end{aligned}$$

Where the last line follows from $d \ge 2$ and $p = O(\log d)$, so that $1 - \frac{1}{d} \ge \frac{1}{2}$ and $d^{3/p} = O(1)$. We conclude that $\|h\|_p = \Omega(1) = \Omega(1)\|h\|_\infty$, and therefore that $\|\mathcal{P}\mathbf{x}\|_p \approx_C \|\mathcal{P}\mathbf{x}\|_\infty$.

Then, we finally combine this with the subspace embedding from the prior lemma to get

$$\frac{1}{C}\|\mathcal{P}\mathbf{x}\|_\infty \le \|\boldsymbol{R}\boldsymbol{A}\mathbf{x}\|_\infty \le C\|\mathcal{P}\mathbf{x}\|_\infty$$

∎

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Now that we have a subspace embedding, we can complete the proof that Algorithm 5 is correct.

Proof. Let $\mathbf{x}^* := \operatorname{argmin}_{\mathbf{x}} \|\mathcal{P}\mathbf{x} - f\|_\infty$ a true optimal solution. We first bound $\|\boldsymbol{RA}\mathbf{x}^* - \boldsymbol{R}\mathbf{b}\|_\infty \le C\|\mathcal{P}\mathbf{x}^* - f\|_\infty$:

$$\begin{aligned}
[R]_{ii} &= (\tfrac{dp}{n}\sqrt{1-t_i^2})^{1/O(\log d)} \\
&\le (\Theta(\tfrac{d\cdot\log d}{d\operatorname{polylog} d}))^{1/O(\log d)} \\
&= (\Theta(\tfrac{1}{\operatorname{polylog} d}))^{1/O(\log d)} \\
&= O(1) \\
\|\boldsymbol{RA}\mathbf{x}^* - \boldsymbol{R}\mathbf{b}\|_\infty &= \sup_{i\in[n]} \left| R_{ii}\left[\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\right]_i \right| \\
&= \sup_{i\in[n]} \left| R_{ii}\left([\mathcal{P}\mathbf{x}^* - f](t_i)\right) \right| \\
&\le \sup_{t\in[-1,1]} O(1) \left| [\mathcal{P}\mathbf{x}^*](t_i) - f(t_i) \right| \\
&= O(1)\|\mathcal{P}\mathbf{x}^* - f\|_\infty \qquad (6.9)
\end{aligned}$$

And this bound suffices to prove our guarantee. Let $\hat{\mathbf{x}} := \operatorname{argmin}_{\mathbf{x}} \|\boldsymbol{RA}\mathbf{x} - \boldsymbol{R}\mathbf{b}\|_\infty$ be the solution returned in line 4 of Algorithm 5. Then,

$$\begin{aligned}
\|\mathcal{P}\hat{\mathbf{x}} - f\|_\infty &\le \|\mathcal{P}\hat{\mathbf{x}} - \mathcal{P}\mathbf{x}^*\|_\infty + \|\mathcal{P}\mathbf{x}^* - f\|_\infty \\
\text{(Subspace Embedding)} \quad &\le C\|\boldsymbol{RA}\hat{\mathbf{x}} - \boldsymbol{RA}\mathbf{x}^*\|_\infty + \|\mathcal{P}\mathbf{x}^* - f\|_\infty \\
&\le C(\|\boldsymbol{RA}\hat{\mathbf{x}} - \boldsymbol{R}\mathbf{b}\|_\infty + \|\boldsymbol{RA}\mathbf{x}^* - \boldsymbol{R}\mathbf{b}\|_\infty) + \|\mathcal{P}\mathbf{x}^* - f\|_\infty \\
\text{(Optimality of } \hat{\mathbf{x}}) \quad &\le 2C\|\boldsymbol{RA}\mathbf{x}^* - \boldsymbol{R}\mathbf{b}\|_\infty + \|\mathcal{P}\mathbf{x}^* - f\|_\infty \\
\text{(Equation 6.9)} \quad &\le O(1)\|\mathcal{P}\mathbf{x}^* - f\|_\infty + \|\mathcal{P}\mathbf{x}^* - f\|_\infty \\
&= O(1)\|\mathcal{P}\mathbf{x}^* - f\|_\infty
\end{aligned}$$

Which completes the proof. ∎

7 Analysis of the Clipped Chebyshev Measure

As mentioned in Section 4, the Chebyshev measure itself is not sufficient to achieve the approximate Lewis weight property for $\mathcal{P}$, since the Chebyshev measure grows to infinity as $|t| \to 1$ while the leverage function is bounded. Thus we instead analyze the following clipped measure: $w(t) := \min\{C(d+1)^2, v(t)\} = \min\{C(d+1)^2, \frac{d+1}{\pi\sqrt{1-t^2}}\}$ and prove the following result:

THEOREM 2.2 RESTATED. *There are fixed constants c_1, c_2 such that, for all $p \in [\frac{2}{3}, 2]$ and $t \in [-1, 1]$,*

$$\frac{c_1}{\log^3 d} \le \frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} \le c_2.$$

The basic flow of the proof is broken into two portions. First, recall the overall shape of the rescaled leverage function:

$$\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) = \max_{\mathbf{x}\in\mathbb{R}^{d+1}} \frac{([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2} = (w(t))^{1-\frac{2}{p}} \max_{q:\deg(q)\le d} \frac{(q(t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}q\|_2^2} \qquad (7.10)$$

We need to show that this leverage function is close to $w(t)$ for all $t \in [-1, 1]$. We split this analysis into two parts:

1. **The "Middle Region"** with $w(t) = v(t)$, so that $|t| \le 1 - O(\frac{1}{d^2})$:
   We show in Section 7.1 that $\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2$ and $\|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2$ are similar enough that $\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}] \approx \tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}]$ in this region, and so the analysis of Theorem 2.2 is tight enough to ensure the almost Lewis weight property here.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

2. **The "Endcap Region"** with $w(t) = C(d+1)^2$, so that $|t| \geq 1 - O(\frac{1}{d^2})$:
We know that $w(t)$ and $v(t)$ are very different here, so we use the fact that $w(t) = C(d+1)^2$ is independent of t. This endcap analysis also proceeds in two steps:

   - **Upper bound $\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) \leq O(w(t)) = O(d^2)$:**
   In Section 7.2.1, we note that $w(t) \leq C(d+1)^2$ for all $t \in [-1,1]$. We use this to lower bound $\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 \geq \frac{1}{C(d+1)^2}\|\mathcal{P}\mathbf{x}\|_2^2$, and reduce the second form in Equation 7.10 to the unweighted leverage function for $\mathcal{P}$. We appeal to our earlier bound on the leverage function for $\mathcal{P}$ from Section 4.1.
   - **Lower bound $\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) \leq \Omega(w(t)\log^3(d)) = \Omega(d^2\log^3(d))$:**
   In Section 7.2.2, we plug in a spike polynomial that approximates $t \mapsto t^{d^2}$ into the rightmost term in Equation 7.10, and evaluate the numerator and denominator for that polynomial.

We again break up the analysis into the slightly more approachable $p = 1$ setting and the more complete $p \in [\frac{2}{3}, 2]$ setting. Additionally, in this section, we refer to the middle region as

$$\mathcal{I}_{mid} := \{t \mid w(t) = v(t)\} = \left[\sqrt{1 - \tfrac{1}{\pi^2(d+1)^2C^2}}, \sqrt{1 + \tfrac{1}{\pi^2(d+1)^2C^2}}\right]$$

and the endcap region as $\mathcal{I}_{cap} := [-1,1] \setminus \mathcal{I}_{mid}$. We also often use the notation $x \approx_\alpha y$ with $\alpha \geq 1$ to mean that $\frac{1}{\alpha}y \leq x \leq \alpha y$. Lastly, to reduce the messiness of the analysis, we omit the change-of-basis matrix that was used in prior sections $\boldsymbol{U}$. For $p = 1$ analysis, Chebyshev polynomials of the second kind are used. For $p = 2$ analysis, Legendre polynomials are used. For $p \in (\frac{2}{3}, 2)$ analysis, Ultraspherical (i.e. Jacobi) polynomials are used.

As an aside, when $p > 2$ this analysis breaks down in a few places since $\frac{1}{2} - \frac{1}{p}$ swaps from being negative to positive. For instance, this means that $t \mapsto (w(t))^{\frac{1}{2}-\frac{1}{p}}$ is maximized in the middle region for $p < 2$ but is maximized in the endcap for $p > 2$.

7.1 Middle Region Analysis for $p = 1$ Our main goal in this section is to prove Lemma 7.4, which states that $\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} = \Theta(1)$. We first recall our bound on the leverage function in Section 4.1:

LEMMA 7.1. *The leverage function for $\mathcal{P}$ has $\tau[\mathcal{P}](t) \leq \frac{(d+1)^2}{2}$ for all $t \in [-1,1]$.*

We use this lemma to (1) analyze the behavior of $\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)}$ on $\mathcal{I}_{mid}$ by showing that the leverage functions on the operators $\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)$ and $\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t)$ are very similar inside this middle region $\mathcal{I}_{mid}$ in Section 7.1.1 and (2) upper and lower bound the ratio of $\frac{\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t)}{v(t)}$ in Section 7.1.2. Using these bounds, we then prove Lemma 7.4 in Section 7.1.3.

7.1.1 Relating $\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}]$ to $\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}]$ In this section, our main goal is to show in Corollary 7.1 that $\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} \approx_{\frac{2}{\pi^2C^2}} \frac{\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t)}{v(t)}$ for $t \in \mathcal{I}_{mid}$, where we recall that $\mathcal{I}_{mid}$ is defined by

$$\mathcal{I}_{mid} := \{t \mid w(t) = v(t)\} = \left[\sqrt{1 - \tfrac{1}{\pi^2(d+1)^2C^2}}, \sqrt{1 + \tfrac{1}{\pi^2(d+1)^2C^2}}\right],$$

so that $w(t) = v(t)$ for $t \in \mathcal{I}_{mid}$. To this end, we first remark that it suffices to show that $\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 \approx_{\frac{2}{\pi^2C^2}} \|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2$. To see why this suffices, consider the definitions of the leverage functions:

$$\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t) = \max_{\mathbf{x}} \frac{([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2} = \max_{\mathbf{x}} \frac{([\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2} \approx_{\frac{2}{\pi^2C^2}} \max_{\mathbf{x}} \frac{([\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2} = \tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t).$$

Hence, we first show in Lemma 7.2 that $\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 \approx_{\frac{\pi^2C^2}{\pi^2C^2-1}} \|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2$.

LEMMA 7.2. *For all $\mathbf{x} \in \mathbb{R}^{d+1}$, we have*

$$\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 \approx_{\frac{\pi^2C^2}{\pi^2C^2-1}} \|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Proof. We start by looking at the difference between $\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2$ and $\|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2$.

$$\begin{aligned}\left|\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 - \|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2\right| &= \left|\int_{-1}^{1}([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2 - ([\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2\ dt\right| \\ &= \left|\int_{\mathcal{I}_{cap}}([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2 - ([\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2\ dt + \int_{\mathcal{I}_{mid}}([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2 - ([\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2\ dt\right|.\end{aligned}$$

Since $[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t) = [\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t)$ for $t \in \mathcal{I}_{mid}$, then $\int_{\mathcal{I}_{mid}}([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2 - ([\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2\ dt = 0$. Moreover, since $w(t)$ is the clipped Chebyshev measure, we have that $w(t) \le v(t)$ and thus $(w(t))^{-\frac{1}{2}} \ge (v(t))^{-\frac{1}{2}}$. Hence,

$$\begin{aligned}\left|\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 - \|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2\right| &= \left|\int_{\mathcal{I}_{cap}}([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2 - ([\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2\ dt\right| \\ &\le \left|\int_{\mathcal{I}_{cap}}([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2\ dt\right| = \int_{\mathcal{I}_{cap}}([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2\ dt.\end{aligned}$$

Because $(w(t))^{-1} = \frac{1}{C(d+1)^2}$ on $\mathcal{I}_{cap}$, then

$$\left|\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 - \|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2\right| \le \frac{1}{C(d+1)^2}\int_{\mathcal{I}_{cap}}([\mathcal{P}\mathbf{x}](t))^2 dt$$

Since Lemma 7.1 implies $([\mathcal{P}\mathbf{x}](t))^2 \le \frac{(d+1)^2}{2}\|\mathcal{P}\mathbf{x}\|_2^2$, then

$$\left|\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 - \|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2\right| \le \frac{1}{C(d+1)^2}\cdot\frac{(d+1)^2}{2}\|\mathcal{P}\mathbf{x}\|_2^2\int_{\mathcal{I}_{cap}} dt = \frac{\|\mathcal{P}\mathbf{x}\|_2^2}{2C}\int_{\mathcal{I}_{cap}} dt.$$

To upper bound the length of the interval $\mathcal{I}_{cap}$, note that $1 - \sqrt{1-\frac{1}{x^2}} \le \frac{1}{x^2}$ for $x^2 \ge 1$. Hence,

$$\textstyle\int_{\mathcal{I}_{cap}} dt = 2\cdot\left(1 - \sqrt{1 - \frac{1}{\pi^2(d+1)^2C^2}}\right) \le \frac{2}{\pi^2(d+1)^2C^2},$$

so that

$$\left|\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 - \|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2\right| \le \frac{\|\mathcal{P}\mathbf{x}\|_2^2}{2C}\cdot\frac{2}{\pi^2(d+1)^2C^2} = \frac{1}{\pi^2(d+1)^2C^3}\|\mathcal{P}\mathbf{x}\|_2^2.$$

Next, we bound the norm $\|\mathcal{P}\mathbf{x}\|_2^2$ using the fact that $w(t) \le C(d+1)^2$ to say that $1 \le \sqrt{C}(d+1)\cdot(w(t))^{-\frac{1}{2}}$, so that

$$\begin{aligned}\|\mathcal{P}\mathbf{x}\|_2^2 &= \int_{-1}^{1}(1\cdot[\mathcal{P}\mathbf{x}](t))^2 dt \\ &\le \int_{-1}^{1}\left(\sqrt{C}(d+1)\cdot(w(t))^{-\frac{1}{2}}\cdot[\mathcal{P}\mathbf{x}](t)\right)^2 dt \\ &= C(d+1)^2\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2.\end{aligned}$$

Therefore,

$$\left|\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 - \|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2\right| \le \frac{C(d+1)^2}{\pi^2(d+1)^2C^3}\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 = \frac{1}{\pi^2C^2}\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2.$$

Rearranging this inequality,

$$\left|1 - \frac{\|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2}\right| \le \frac{1}{\pi^2C^2}$$

or equivalently,

$$1 - \frac{1}{\pi^2C^2} \le \frac{\|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2} \le 1 + \frac{1}{\pi^2C^2} \le \frac{1}{1-\frac{1}{\pi^2C^2}},$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

for $C > \frac{1}{\pi}$. Since $\frac{1}{1-\frac{1}{\pi^2C^2}} = \frac{\pi^2C^2}{\pi^2C^2-1}$, then we have the multiplicative error guarantee

$$\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 \approx_{\frac{\pi^2C^2}{\pi^2C^2-1}} \|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2$$

for $C > \frac{1}{\pi} \approx 0.312$. ∎

We now complete the formal proof of Corollary 7.1.

COROLLARY 7.1. $\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} \approx_{\frac{\pi^2C^2}{\pi^2C^2-1}} \frac{\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t)}{v(t)}$ *for* $t \in \mathcal{I}_{mid}$.

Proof. By Lemma 7.2, we have that $\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 \approx_{\frac{\pi^2C^2}{\pi^2C^2-1}} \|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2$ for all $\mathbf{x} \in \mathbb{R}^{d+1}$. Since $w(t) = v(t)$ for $t \in \mathcal{I}_{mid}$, Lemma 7.2 implies through the definition of the leverage functions that

$$\begin{aligned}
\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} &= \frac{1}{w(t)} \max_{\mathbf{x}} \frac{([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2} \\
&= \frac{1}{v(t)} \max_{\mathbf{x}} \frac{([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2} \\
&\approx_{\frac{\pi^2C^2}{\pi^2C^2-1}} \frac{1}{v(t)} \max_{\mathbf{x}} \frac{([\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{V}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2} \\
&= \frac{\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t)}{v(t)},
\end{aligned}$$

as desired. ∎

7.1.2 Relating $\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}]$ to $v(t)$ In this section, we relate $\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}]$ to $v(t)$ for $t \in \mathcal{I}_{mid}$, which will ultimately allow us to relate $\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}]$ to $w(t)$ in Section 7.1.3, using Corollary 7.1.

LEMMA 7.3. *For* $t \in \mathcal{I}_{mid}$, *we have that* $\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t) \approx_{(\frac{5}{4}+\frac{\pi C}{2})} v(t)$.

Proof. Note that the claim is equivalent to the statement that $\frac{\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t)}{v(t)} \in (\frac{1}{\gamma}, \gamma)$ for $\gamma \leq \frac{5}{4} + \frac{\pi C}{2}$. We will use the relationship $\frac{-1}{\sqrt{1-t^2}} \leq U_i(t) \leq \frac{1}{\sqrt{1-t^2}}$ to prove this.

Specifically, we ensure the two traits

$$\begin{aligned}
\frac{\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t)}{v(t)} &\leq 1 + \frac{1 + \frac{1}{\sqrt{1-t^2}}}{2(d+1)} \leq \gamma \\
\frac{\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t)}{v(t)} &\geq 1 + \frac{1 - \frac{1}{\sqrt{1-t^2}}}{2(d+1)} \geq \frac{1}{\gamma}.
\end{aligned}$$

Solving these two inequalities on the right hand side yields

$$|t| \leq \sqrt{1 - \frac{1}{(2(d+1)(\gamma-1)-1)^2}} \qquad \text{and} \qquad |t| \leq \sqrt{1 - \frac{1}{(2(d+1)(\frac{1}{\gamma}-1)-1)^2}},$$

respectively. Observe that the guarantee on the left implies the guarantee on the right, so we just ensure that one trait. Rather, we should think of

$$\mathcal{I}_\gamma := \left\{ t \;\middle|\; |t| \leq \sqrt{1 - \frac{1}{(2(d+1)(\gamma-1)-1)^2}} \right\}$$

as the set of time points where we have $\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t) \approx_\gamma v(t)$. We now ensure that this interval $\mathcal{I}_\gamma$ entirely contains the middle region $\mathcal{I}_{mid}$, i.e., $\mathcal{I}_{mid} \subset \mathcal{I}_\gamma$. Note that $t \in \mathcal{I}_{mid}$ implies that

$$t \leq \sqrt{1 - \frac{1}{\pi^2(d+1)^2C^2}}.$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

For $\gamma = 1 + \frac{\pi}{2}C + \frac{1}{2(d+1)}$, note that we have

$$t \le \sqrt{1 - \frac{1}{(2(d+1)(\gamma-1)-1)^2}},$$

as desired. Hence, $\mathcal{I}_{mid} = \mathcal{I}_\gamma$ for $\pi^2(d+1)^2C^2 = (2(d+1)(\gamma-1)-1)^2$ or equivalently, $\gamma = 1 + \frac{\pi}{2}C + \frac{1}{2(d+1)}$. Since $d \ge 1$ implies

$$1 + \frac{\pi}{2}C + \frac{1}{2(d+1)} \le 1 + \frac{\pi}{2}C + \frac{1}{4} = \frac{5}{4} + \frac{\pi}{2}C,$$

then $\mathcal{I}_{mid} \subset \mathcal{I}_\gamma$ for $\gamma \le \frac{5}{4} + \frac{\pi}{2}C$. Therefore, the set $\mathcal{I}_\gamma$ where the leverage scores of $\mathcal{V}^{-\frac{1}{2}}\mathcal{P}$ are γ-close to $v(t)$ covers the set of time-samples *not* in the cap for $\gamma \le \frac{5}{4} + \frac{\pi}{2}C$. Equivalently, we have that $\tau[\mathcal{V}^{-\frac{1}{2}}\mathcal{P}](t) \approx_{(\frac{5}{4}+\frac{\pi C}{2})} v(t)$ for $t \in \mathcal{I}_{mid}$. ∎

7.1.3 Complete Result in the Middle We now finally relate $\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)}$ by using Corollary 7.1 and Lemma 7.3.

LEMMA 7.4. *For $t \in \mathcal{I}_{mid}$, we have*

$$\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} = \Theta(1).$$

Proof. By Corollary 7.1 and Lemma 7.3, we have that for $t \in \mathcal{I}_{mid}$,

$$\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t) \approx_\alpha v(t)$$

where $\alpha = \frac{\pi^2C^2}{\pi^2C^2-1} \cdot \left(\frac{5}{4} + \frac{\pi C}{2}\right)$ for some constant $C > \frac{1}{\pi} \approx 0.312$. Furthermore, since $v(t) = w(t)$ in the region $t \in \mathcal{I}_{mid}$, this further implies $\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t) \approx_\alpha w(t)$, as desired. ∎

7.2 Endcap Region Analysis for $p = 1$ We now turn to $t \in \mathcal{I}_{cap}$, and we will show that

$$\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t) \approx_{\frac{1}{2C}} w(t)$$

for $t \in \mathcal{I}_{cap}$. Thus it suffices to upper and lower bound the ratio $\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)}$.

7.2.1 Upper Bounding the Ratio. In this section, we provide an upper bound on the ratio $\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)}$. Namely, we show in Lemma 7.5 that there exists an absolute constant C, the same constant $C > \frac{1}{\pi}$ in the definition of the clipped Chebyshev measure, such that $\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} \le \frac{1}{2C}$ for all $t \in \mathcal{I}_{cap}$.

LEMMA 7.5. *For $t \in \mathcal{I}_{cap}$, we have*

$$\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} = O(1).$$

Proof. Since $\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t) = \max_{\mathbf{x}} \frac{([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2}$ and $w(t) \le C(d+1)^2$ for all $t \in [-1,1]$, we first lower bound $\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2$ by

$$\begin{aligned}\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2 &= \int_{-1}^{1} \frac{1}{w(t)}([\mathcal{P}\mathbf{x}](t))^2 \, dt \\ &\ge \int_{-1}^{1} \frac{1}{C(d+1)^2}([\mathcal{P}\mathbf{x}](t))^2 \, dt = \frac{1}{C(d+1)^2}\|\mathcal{P}\mathbf{x}\|_2^2.\end{aligned}$$

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Then we can directly tackle the leverage function:

$$\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t) = \max_{\mathbf{x}} \frac{([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2} = \frac{1}{w(t)} \max_{\mathbf{x}} \frac{([\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2}$$
$$= \frac{1}{C(d+1)^2} \max_{\mathbf{x}} \frac{([\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2}$$

since $w(t) = C(d+1)^2$ for $t \in \mathcal{I}_{cap}$. Thus,

$$\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t) \leq \frac{1}{C(d+1)^2} \max_{\mathbf{x}} \frac{([\mathcal{P}\mathbf{x}](t))^2}{\frac{1}{C(d+1)^2}\|\mathcal{P}\mathbf{x}\|_2^2} = \tau[\mathcal{P}](t) \leq \frac{(d+1)^2}{2}.$$

Then we can then conclude

$$\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} \leq \frac{\frac{(d+1)^2}{2}}{C(d+1)^2} = \frac{1}{2C}.$$

□

7.2.2 Lower Bounding the Ratio. In this section, we provide a lower bound on the ratio $\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)}$. Namely, we show in Lemma 7.6 that there exists an absolute constant C' such that $\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} \geq \frac{C'}{\log^3 d}$ for all $t \in \mathcal{I}_{cap}$. We first require the following structural result from polynomial approximation theory.

THEOREM 7.1. (LOW-DEGREE APPROXIMATION OF HIGH-DEGREE POLYNOMIAL, THEOREM 3.3 IN [SV14]) *For any positive integers s and d, there exists a degree d polynomial F such that*

$$\sup_{t\in[-1,1]} |f(t) - t^s| \leq 2e^{-\frac{d^2}{s}}.$$

Moreover, for any $\delta > 0$ and $d \geq \left\lceil \sqrt{2s \log \frac{2}{\delta}} \right\rceil$, there exists a polynomial f of degree d such that

$$\sup_{t\in[-1,1]} |f(t) - t^s| \leq \delta.$$

LEMMA 7.6. *For $t \in \mathcal{I}_{cap}$, we have*

$$\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} = \Omega\left(\frac{1}{\log^3 d}\right).$$

Proof. To lower bound the ratio $\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)}$, we first note that for $t \in \mathcal{I}_{cap}$, we have that $w(t) = C(d+1)^2$ and thus it suffices to lower bound

$$\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t) = \max_{\mathbf{x}} \frac{([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2}$$

by analyzing the quantity for a specific choice of $\mathbf{x} \in \mathbb{R}^{d+1}$.

Let $q = O\left(\frac{(d+1)^2}{\log d}\right)$ so that by Theorem 7.1, there exists a degree d polynomial f such that

$$\sup_{t\in[-1,1]} |f(t) - t^q| \leq d^{-\gamma},$$

for some constant $\gamma > 0$. We set $\mathbf{x} \in \mathbb{R}^{d+1}$ so that the operator $\mathcal{P}\mathbf{x}$ corresponds to $f(t)$ and lower bound $\frac{([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2}$.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

First, note that since $t^q = 1$ at $t = 1$, then we have $f(1) \geq 1 - d^{-\gamma}$. Similarly, since $|t| \geq \sqrt{1 - \frac{1}{\pi^2(d+1)^2C^2}} \geq 1 - \frac{1}{2\pi^2(d+1)^2C^2}$ for $t \in \mathcal{I}_{cap}$, then we have $t^q \geq \frac{1}{4}$ since $q = O\left(\frac{(d+1)^2}{\log d}\right)$. Thus, we have $f(t) \geq \frac{1}{4} - d^{-\gamma}$ for all $t \in \mathcal{I}_{cap}$. Since $w(t) = C(d+1)^2$ for all $t \in \mathcal{I}_{cap}$, then

$$([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2 \geq \frac{1}{8C(d+1)^2}.$$

It remains to upper bound $\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2$ when the operator $\mathcal{P}\mathbf{x}$ corresponds to $F(t)$. Since $\sup_{t\in[-1,1]} |f(t) - t^q| \leq d^{-\gamma}$, then we have

$$\begin{aligned}\|\mathcal{W}^{-\frac{1}{2}}f\|_2^2 &= \int_{-1}^{1} \frac{1}{w(t)}(f(t))^2\,dt \\ &\leq 2\int_{-1}^{1} \frac{1}{w(t)}d^{-2\gamma}\,dt + 2\int_{-1}^{1} \frac{1}{w(t)}t^{2q}\,dt.\end{aligned}$$

Since $w(t) = \min\{C(d+1)^2,\ \frac{d+1}{\pi\sqrt{1-t^2}}\}$, then $\frac{1}{w(t)} \leq \frac{\pi}{d+1}$. Thus,

$$\|\mathcal{W}^{-\frac{1}{2}}f\|_2^2 \leq \frac{4\pi d^{-2\gamma}}{d+1} + 4\int_{0}^{1} \frac{1}{w(t)}t^{2q}\,dt.$$

We decompose the interval $[0,1]$ into $\mathcal{I}_1 = \left[0, \sqrt{1 - \frac{C^2\pi^2\log^2 d}{(d+1)^2}}\right)$ and $\mathcal{I}_2 = \left[\sqrt{1 - \frac{C^2\pi^2\log^2 d}{(d+1)^2}}, 1\right]$. Note that for $t \in \mathcal{I}_1$, we have $t \leq 1 - \frac{C^2\pi^2\log^2 d}{2(d+1)^2}$ and thus $t^{2q} \leq \exp\left(-O\left(C^2\pi^2\log d\right)\right)$ for $q = O\left(\frac{(d+1)^2}{\log d}\right)$. Hence for sufficiently large $C > 0$, we have that $t^{2q} \leq \frac{1}{16\pi(d+1)^3}$ for all $t \in \mathcal{I}_1$. Thus since $\frac{1}{w(t)} \leq \frac{\pi}{d+1}$, then

$$4\int_{\mathcal{I}_1} \frac{1}{w(t)}t^{2q}\,dt \leq \frac{16\pi}{d+1}\int_{\mathcal{I}_1} t^{2q}\,dt \leq \frac{1}{(d+1)^4}.$$

Note that $|\mathcal{I}_2| \leq \frac{C^2\pi^2\log^2 d}{2(d+1)^2}$ and $t^{2q} \leq 1$ for $t \in \mathcal{I}_2$. Moreover for $t \in \mathcal{I}_2$, we have $\frac{d+1}{\pi\sqrt{1-t^2}} \geq \frac{C(d+1)^2}{\log d}$ so that $\frac{1}{w(t)} \leq \frac{\log d}{C(d+1)^2}$. Hence,

$$\begin{aligned}\int_{\mathcal{I}_2} \frac{1}{w(t)}t^{2q}\,dt &\leq \int_{\mathcal{I}_2} \frac{\log d}{C(d+1)^2}\,dt \\ &\leq \frac{\log d}{C(d+1)^2}\cdot\frac{C^2\pi^2\log^2 d}{2(d+1)^2} = \frac{C\pi^2\log^3 d}{2(d+1)^4}.\end{aligned}$$

Therefore in summary, we have

$$\begin{aligned}\|\mathcal{W}^{-\frac{1}{2}}f\|_2^2 &\leq \frac{4\pi d^{-2\gamma}}{d+1} + 4\int_{0}^{1} \frac{1}{w(t)}t^{2q}\,dt \\ &= \frac{4\pi d^{-2\gamma}}{d+1} + 4\int_{\mathcal{I}_1} \frac{1}{w(t)}t^{2q}\,dt + 4\int_{\mathcal{I}_2} \frac{1}{w(t)}t^{2q}\,dt \\ &\leq \frac{4\pi d^{-2\gamma}}{d+1} + \frac{1}{(d+1)^4} + \frac{C\pi^2\log^3 d}{2(d+1)^4}.\end{aligned}$$

Hence for sufficiently large $\gamma > 0$, we have that

$$\|\mathcal{W}^{-\frac{1}{2}}f\|_2^2 = O\left(\frac{\log^3 d}{d^4}\right).$$

Combined with the previous bound of $([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2 \geq \frac{1}{8C(d+1)^2} = \Omega\left(\frac{1}{d^2}\right)$, then

$$\frac{([\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{-\frac{1}{2}}\mathcal{P}\mathbf{x}\|_2^2} = \Omega\left(\frac{d^2}{\log^3 d}\right).$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Finally, since $w(t) \le C(d+1)^2$, then

$$\frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} = \Omega\left(\frac{1}{\log^3 d}\right).$$

∎

7.3 Putting It All Together We finally obtain Theorem 2.2 from Lemma 7.4, Lemma 7.5, and Lemma 7.6.

THEOREM 2.2 RESTATED. *There are fixed constants c_1, c_2, c_3 such that, letting $w(t) = \min\left(c_1(d+1)^2, \frac{d+1}{\pi\sqrt{1-t^2}}\right)$ be the clipped Chebyshev measure on $[-1,1]$ and letting $\mathcal{W}$ be the corresponding diagonal operator with $[\mathcal{W}x](t) = w(t)\cdot x(t)$, for any $t \in [-1,1]$,*

$$\frac{c_2}{\log^3 d} \le \frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} \le c_3.$$

Proof. We consider casework on $t \in [-1,1]$. Recall that

$$\mathcal{I}_{mid} := \{t \mid w(t) = v(t)\} = \left[\sqrt{1 - \tfrac{1}{\pi^2(d+1)^2C^2}}, \sqrt{1 + \tfrac{1}{\pi^2(d+1)^2C^2}}\right]$$

and $\mathcal{I}_{cap} := [-1,1] \setminus \mathcal{I}_{mid}$. We have from Lemma 7.4 that there exists a constant $C_0 \ge 1$ such that $\frac{1}{C_0} \le \frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} \le C_0$ for all $t \in \mathcal{I}_{mid}$. We have from Lemma 7.5 and Lemma 7.6 that there exist constants C_3, C_4 such that

$$\frac{C_3}{\log^3 d} \le \frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} \le C_4$$

for all $t \in \mathcal{I}_{cap}$. Thus by setting $C_1 = \min\left(C_3, \frac{1}{C_0}\right)$ and $C_2 = \max(C_0, C_4)$, we have that

$$\frac{C_1}{\log^3 d} \le \frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} \le C_2$$

for all $t \in [-1,1]$. ∎

We now move onto the slightly messier analysis which works for all $p \in [1,2]$. The core ideas are all the same, but the mathematical arguments are slightly more nuanced.

7.4 Middle Region Analysis for $p \in [\frac{2}{3}, 2]$ In this section, we show that $\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} = \Theta(1)$ for $t \in \mathcal{I}_{mid}$ defined by

$$\mathcal{I}_{mid} := \{t \mid w(t) = v(t)\} = \left[\sqrt{1 - \tfrac{1}{\pi^2(d+1)^2C^2}}, \sqrt{1 + \tfrac{1}{\pi^2(d+1)^2C^2}}\right]$$

and the clipped Chebyshev measure $w(t)$ defined by

$$w(t) := \min\{C(d+1)^2,\ v(t)\} = \min\{C(d+1)^2,\ \tfrac{d+1}{\pi\sqrt{1-t^2}}\}.$$

7.4.1 Relating $\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}]$ to $\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}]$ We first show that $\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} \approx_{\frac{2}{\pi^2C^2}} \frac{\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{v(t)}$ for $t \in \mathcal{I}_{mid}$. Observe that since $\mathcal{I}_{mid}$ is defined by

$$\mathcal{I}_{mid} := \{t \mid w(t) = v(t)\} = \left[\sqrt{1 - \tfrac{1}{\pi^2(d+1)^2C^2}}, \sqrt{1 + \tfrac{1}{\pi^2(d+1)^2C^2}}\right],$$

then we have $w(t) = v(t)$ for $t \in \mathcal{I}_{mid}$. Thus it suffices to show that $\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 \approx_{\frac{2}{\pi^2C^2}} \|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2$ since

$$\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) = \max_{\mathbf{x}} \frac{([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2} = \max_{\mathbf{x}} \frac{([\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2}$$
$$\approx_{\frac{2}{\pi^2C^2}} \max_{\mathbf{x}} \frac{([\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2} = \tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t).$$

Therefore, we first show that $\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 \approx_{\frac{2}{\pi^2C^2}} \|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2$.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

LEMMA 7.7. *For all* $\mathbf{x} \in \mathbb{R}^{d+1}$, *we have*

$$\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 \approx_{\frac{\pi^2C^2}{\pi^2C^2-1}} \|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2$$

Proof. We first bound the difference between $\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2$ and $\|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2$.

$$\begin{aligned}\left|\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 - \|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2\right| &= \left|\int_{-1}^{1}([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2 - ([\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2\ dt\right| \\ &= \left|\int_{\mathcal{I}_{cap}}([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2 - ([\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2\ dt + \int_{\mathcal{I}_{mid}}([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2 - ([\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2\ dt\right|.\end{aligned}$$

Because $[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t) = [\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t)$ for $t \in \mathcal{I}_{mid}$, then it follows that $\int_{\mathcal{I}_{mid}}([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2 - ([\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2\ dt = 0$. Since $w(t)$ is the clipped Chebyshev measure, we have that $w(t) \le v(t)$ and thus $(w(t))^{\frac{1}{2}-\frac{1}{p}} \ge (v(t))^{\frac{1}{2}-\frac{1}{p}}$. Therefore,

$$\begin{aligned}\left|\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 - \|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2\right| &= \left|\int_{\mathcal{I}_{cap}}([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2 - ([\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2\ dt\right| \\ &\le \left|\int_{\mathcal{I}_{cap}}([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2\ dt\right| = \int_{\mathcal{I}_{cap}}([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2\ dt.\end{aligned}$$

Since $w(t) = C(d+1)^2$ on $\mathcal{I}_{cap}$,

$$\left|\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 - \|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2\right| \le (C(d+1)^2)^{1-\frac{2}{p}}\int_{\mathcal{I}_{cap}}([\mathcal{P}\mathbf{x}](t))^2 dt.$$

By Lemma 7.1, we have that $([\mathcal{P}\mathbf{x}](t))^2 \le \frac{(d+1)^2}{2}\|\mathcal{P}\mathbf{x}\|_2^2$. Thus,

$$\begin{aligned}\left|\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 - \|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2\right| &\le (C(d+1)^2)^{1-\frac{2}{p}} \cdot \frac{(d+1)^2}{2}\|\mathcal{P}\mathbf{x}\|_2^2\int_{\mathcal{I}_{cap}} dt \\ &= \frac{C^{1-\frac{2}{p}}(d+1)^{4-\frac{4}{p}}}{2}\|\mathcal{P}\mathbf{x}\|_2^2\int_{\mathcal{I}_{cap}} dt.\end{aligned}$$

We upper bound the length of the interval $\mathcal{I}_{cap}$ by observing that $1 - \sqrt{1-\frac{1}{x^2}} \le \frac{1}{x^2}$ for $x^2 \ge 1$ and thus,

$$\int_{\mathcal{I}_{cap}} dt = 2 \cdot \left(1 - \sqrt{1 - \frac{1}{\pi^2(d+1)^2C^2}}\right) \le \frac{2}{\pi^2(d+1)^2C^2}.$$

Therefore,

$$\left|\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 - \|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2\right| \le \frac{2}{\pi^2(d+1)^2C^2} \cdot \frac{C^{1-\frac{2}{p}}(d+1)^{4-\frac{4}{p}}}{2}\|\mathcal{P}\mathbf{x}\|_2^2 = \frac{(d+1)^{2-\frac{4}{p}}}{C^{\frac{2}{p}+1}\pi^2} \cdot \|\mathcal{P}\mathbf{x}\|_2^2.$$

We then bound the norm $\|\mathcal{P}\mathbf{x}\|_2^2$ by noting that $w(t) \le C(d+1)^2$. Thus, $1 \le C^{\frac{1}{p}-\frac{1}{2}}(d+1)^{\frac{2}{p}-1} \cdot (w(t))^{\frac{1}{2}-\frac{1}{p}}$, so that

$$\begin{aligned}\|\mathcal{P}\mathbf{x}\|_2^2 &= \int_{-1}^{1}(1 \cdot [\mathcal{P}\mathbf{x}](t))^2 dt \\ &\le \int_{-1}^{1}\left(C^{\frac{1}{p}-\frac{1}{2}}(d+1)^{\frac{2}{p}-1} \cdot (w(t))^{\frac{1}{2}-\frac{1}{p}} \cdot [\mathcal{P}\mathbf{x}](t)\right)^2 dt \\ &= C^{\frac{2}{p}-1}(d+1)^{\frac{4}{p}-2}\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2.\end{aligned}$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Therefore,

$$\left|\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 - \|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2\right| \leq \frac{(d+1)^{2-\frac{4}{p}}}{C^{\frac{2}{p}+1}\pi^2} \cdot C^{\frac{2}{p}-1}(d+1)^{\frac{4}{p}-2}\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 = \tfrac{1}{\pi^2C^2}\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2.$$

Rearranging this inequality, we have that

$$\left|1 - \frac{\|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2}\right| \leq \frac{1}{\pi^2C^2}$$

or equivalently,

$$1 - \frac{1}{\pi^2C^2} \leq \frac{\|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2} \leq 1 + \frac{1}{\pi^2C^2} \leq \frac{1}{1-\frac{1}{\pi^2C^2}},$$

for $C > \frac{1}{\pi}$. Since $\frac{1}{1-\frac{1}{\pi^2C^2}} = \frac{\pi^2C^2}{\pi^2C^2-1}$, then we have the multiplicative error guarantee

$$\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 \approx_{\frac{\pi^2C^2}{\pi^2C^2-1}} \|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2$$

for $C > \frac{1}{\pi} \approx 0.312$. ∎

We now relate $\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)}$ to $\frac{\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{v(t)}$ for $t \in \mathcal{I}_{mid}$

COROLLARY 7.2. $\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} \approx_{\frac{\pi^2C^2}{\pi^2C^2-1}} \frac{\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{v(t)}$ *for* $t \in \mathcal{I}_{mid}$.

Proof. By Lemma 7.7, $\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 \approx_{\frac{\pi^2C^2}{\pi^2C^2-1}} \|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2$ for all $\mathbf{x} \in \mathbb{R}^{d+1}$. Since $w(t) = v(t)$ for $t \in \mathcal{I}_{mid}$, it follows from the definition of the leverage functions that

$$\begin{aligned}
\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} &= \frac{1}{w(t)}\max_{\mathbf{x}}\frac{([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2} \\
&= \frac{1}{v(t)}\max_{\mathbf{x}}\frac{([\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2} \\
&\approx_{\frac{\pi^2C^2}{\pi^2C^2-1}} \frac{1}{v(t)}\max_{\mathbf{x}}\frac{([\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2} \\
&= \frac{\tau[\mathcal{V}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{v(t)},
\end{aligned}$$

as desired. ∎

7.4.2 Complete Result in the Middle We now show that $\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)$ and $w(t)$ are within a constant factor for $t \in \mathcal{I}_{mid}$.

LEMMA 7.8. *For* $t \in \mathcal{I}_{mid}$, *we have*

$$\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} = \Theta(1).$$

Proof. By Corollary 7.2 and Corollary 4.1, we have that for $t \in \mathcal{I}_{mid}$,

$$\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) \approx_\alpha v(t)$$

for $\alpha = \frac{\pi^2C^2}{\pi^2C^2-1} \cdot C_0$ for some constants C_0 and $C > \frac{1}{\pi} \approx 0.312$. Furthermore, since $v(t) = w(t)$ in the region $t \in \mathcal{I}_{mid}$, this further implies $\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) \approx_\alpha w(t)$, as desired. ∎

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

7.5 Endcap Region Analysis for $p \in [\frac{2}{3}, 2]$ We now bound the ratio for $t \in \mathcal{I}_{cap}$, and we will show that

$$\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) \approx_{\frac{1}{2C}} w(t)$$

for $t \in \mathcal{I}_{cap}$. Thus it suffices to upper and lower bound the ratio $\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)}$.

7.5.1 Upper Bounding the Ratio. In this section, we provide an upper bound on the ratio $\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)}$.

LEMMA 7.9. *For $t \in \mathcal{I}_{cap}$, we have*

$$\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} = O(1).$$

Proof. Since $\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) = \max_{\mathbf{x}} \frac{([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2}$ and $w(t) \leq C(d+1)^2$ for all $t \in [-1,1]$, then for $p \in [\frac{2}{3}, 2]$, we can first lower bound $\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2$ by

$$\begin{aligned}\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2 &= \int_{-1}^{1} (w(t))^{1-\frac{2}{p}} ([\mathcal{P}\mathbf{x}](t))^2 \, dt \\ &\geq \int_{-1}^{1} (C(d+1)^2)^{1-\frac{2}{p}} ([\mathcal{P}\mathbf{x}](t))^2 \, dt = (C(d+1)^2)^{1-\frac{2}{p}} \|\mathcal{P}\mathbf{x}\|_2^2.\end{aligned}$$

On the other hand, the leverage function $\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}]$ satisfies

$$\begin{aligned}\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) &= \max_{\mathbf{x}} \frac{([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2} = (w(t))^{1-\frac{2}{p}} \max_{\mathbf{x}} \frac{([\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2} \\ &= (C(d+1)^2)^{1-\frac{2}{p}} \max_{\mathbf{x}} \frac{([\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2}\end{aligned}$$

because $w(t) = C(d+1)^2$ for $t \in \mathcal{I}_{cap}$. Therefore, from the above inequality, we have

$$\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) \leq (C(d+1)^2)^{1-\frac{2}{p}} \max_{\mathbf{x}} \frac{([\mathcal{P}\mathbf{x}](t))^2}{(C(d+1)^2)^{1-\frac{2}{p}}\|\mathcal{P}\mathbf{x}\|_2^2} = \tau[\mathcal{P}](t) \leq \frac{(d+1)^2}{2}.$$

Hence,

$$\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} \leq \frac{\frac{(d+1)^2}{2}}{C(d+1)^2} = \frac{1}{2C},$$

as desired. ∎

7.5.2 Lower Bounding the Ratio. We now lower bound the ratio $\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)}$.

LEMMA 7.10. *For $t \in \mathcal{I}_{cap}$, we have*

$$\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} = \Omega\left(\frac{1}{\log^3 d}\right).$$

Proof. To lower bound the ratio $\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)}$, we observe that $w(t) = C(d+1)^2$ for $t \in \mathcal{I}_{cap}$. Hence, it suffices to lower bound

$$\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t) = \max_{\mathbf{x}} \frac{([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2}$$

by choosing a specific polynomial represented by $\mathbf{x} \in \mathbb{R}^{d+1}$.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

We choose $q = O\left(\frac{(d+1)^2}{\log d}\right)$ so that by Theorem 7.1, there exists a degree d polynomial f such that

$$\sup_{t\in[-1,1]} |f(t) - t^q| \le d^{-\gamma},$$

for some fixed constant $\gamma > 0$ to be set at a later point in the analysis. We choose $\mathbf{x} \in \mathbb{R}^{d+1}$ so that the operator $\mathcal{P}\mathbf{x}$ corresponds to $f(t)$. We then lower bound $\frac{([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2}$.

Because $t^q = 1$ at $t = 1$, then $f(1) \ge 1 - d^{-\gamma}$. Since $|t| \ge \sqrt{1 - \frac{1}{\pi^2(d+1)^2C^2}} \ge 1 - \frac{1}{2\pi^2(d+1)^2C^2}$ for $t \in \mathcal{I}_{cap}$, then $t^q \ge \frac{1}{4}$ for $q = O\left(\frac{(d+1)^2}{\log d}\right)$. Hence, $f(t) \ge \frac{1}{4} - d^{-\gamma}$ for all $t \in \mathcal{I}_{cap}$. Since $w(t) = C(d+1)^2$ for all $t \in \mathcal{I}_{cap}$, then

$$([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2 \ge \frac{1}{8}(C(d+1)^2)^{1-\frac{2}{p}} = \Omega\left(d^{2-\frac{4}{p}}\right).$$

We now upper bound $\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2$ for the operator $\mathcal{P}\mathbf{x}$ that corresponds to $f(t)$. Since $\sup_{t\in[-1,1]} |f(t)-t^q| \le d^{-\gamma}$, then

$$\begin{aligned}\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}f\|_2^2 &= \int_{-1}^{1} (w(t))^{1-\frac{2}{p}}(F(t))^2\,dt \\ &\le 2\int_{-1}^{1}(w(t))^{1-\frac{2}{p}}d^{-2\gamma}\,dt + 2\int_{-1}^{1}(w(t))^{1-\frac{2}{p}}t^{2q}\,dt.\end{aligned}$$

Since $w(t) = \min\{C(d+1)^2,\ \frac{d+1}{\pi\sqrt{1-t^2}}\}$, then $(w(t))^{1-\frac{2}{p}} = O\left(d^{1-\frac{2}{p}}\right)$ for $p \in [\frac{2}{3}, 2]$. Thus,

$$\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}f\|_2^2 \le O\left(d^{1-\frac{2}{p}-2\gamma}\right) + 4\int_0^1 (w(t))^{1-\frac{2}{p}}t^{2q}\,dt.$$

Consider a decomposition of the interval $[0,1]$ into intervals $\mathcal{I}_1 = \left[0, \sqrt{1 - \frac{C^2\pi^2\log^2 d}{(d+1)^2}}\right)$ and $\mathcal{I}_2 = \left[\sqrt{1 - \frac{C^2\pi^2\log^2 d}{(d+1)^2}}, 1\right]$. For $t \in \mathcal{I}_1$, we have $t \le 1 - \frac{C^2\pi^2\log^2 d}{2(d+1)^2}$ so that $t^{2q} \le \exp\left(-O\left(C^2\pi^2\log d\right)\right)$ for $q = O\left(\frac{(d+1)^2}{\log d}\right)$. Thus for sufficiently large $C > 0$, we have that $t^{2q} = O\left(\frac{1}{d^7}\right)$ for all $t \in \mathcal{I}_1$. Because $(w(t))^{1-\frac{2}{p}} \le 1$ for $p \in [\frac{2}{3}, 2]$, then

$$4\int_{\mathcal{I}_1}(w(t))^{1-\frac{2}{p}}t^{2q}\,dt = O\left(\frac{1}{d^7}\right).$$

On the other hand, $|\mathcal{I}_2| \le \frac{C^2\pi^2\log^2 d}{2(d+1)^2}$ and $t^{2q} \le 1$ for $t \in \mathcal{I}_2$. For $t \in \mathcal{I}_2$, we also have either $w(t) = C(d+1)^2$ or $w(t) = \frac{d+1}{\pi\sqrt{1-t^2}} \ge \frac{(d+1)^2}{C\pi\log d}$ so that either way $w(t) \ge \frac{(d+1)^2}{C\pi\log d}$, and so $(w(t))^{1-\frac{2}{p}} = O\left(d^{2-\frac{4}{p}}\log^2 d\right)$. Hence,

$$\begin{aligned}4\int_{\mathcal{I}_2}(w(t))^{1-\frac{2}{p}}t^{2q}\,dt4 &\le \int_{\mathcal{I}_2} O\left(d^{2-\frac{4}{p}}\log^2 d\right)\,dt \\ &\le O\left(d^{2-\frac{4}{p}}\log^2 d\right)\cdot\frac{C^2\pi^2\log^2 d}{2(d+1)^2} = O\left(d^{-\frac{4}{p}}\log^4 d\right).\end{aligned}$$

Thus in all,

$$\begin{aligned}\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}f\|_2^2 &\le O\left(d^{1-\frac{2}{p}-2\gamma}\right) + 4\int_0^1(w(t))^{1-\frac{2}{p}}t^{2q}\,dt \\ &= O\left(d^{1-\frac{2}{p}-2\gamma}\right) + 4\int_{\mathcal{I}_1}(w(t))^{1-\frac{2}{p}}t^{2q}\,dt + 4\int_{\mathcal{I}_2}(w(t))^{1-\frac{2}{p}}t^{2q}\,dt \\ &= O\left(d^{1-\frac{2}{p}-2\gamma}\right) + O\left(\frac{1}{d^7}\right) + O\left(d^{-\frac{4}{p}}\log^4 d\right).\end{aligned}$$

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Hence for $\gamma = 5$, for all $p \in [\frac{2}{3}, 1]$, we have that

$$\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}} F\|_2^2 = O\left(d^{-\frac{4}{p}} \log^4 d\right).$$

Combined with our previous bound that $([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2 \geq \Omega\left(d^{2-\frac{4}{p}}\right)$ for $p \in [\frac{2}{3}, 2]$ and therefore,

$$\frac{([\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}](t))^2}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_2^2} = \Omega\left(\frac{d^2}{\log^4 d}\right).$$

Finally, because $w(t) \leq C(d+1)^2$, then

$$\frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} = \Omega\left(\frac{1}{\log^4 d}\right).$$

□

7.6 Putting It All Together We finally obtain Theorem 2.2 from Lemma 7.8, Lemma 7.9, and Lemma 7.10.

THEOREM 2.2 RESTATED. *There are fixed constants c_1, c_2, c_3 such that, letting $w(t) = \min\left(c_1(d+1)^2, \frac{d+1}{\pi\sqrt{1-t^2}}\right)$ be the clipped Chebyshev measure on $[-1,1]$ and letting $\mathcal{W}$ be the corresponding diagonal operator with $[\mathcal{W}x](t) = w(t) \cdot x(t)$, for any $p \in [\frac{2}{3}, 2]$ and $t \in [-1,1]$,*

$$\frac{c_2}{\log^3 d} \leq \frac{\tau[\mathcal{W}^{-\frac{1}{2}}\mathcal{P}](t)}{w(t)} \leq c_3.$$

Proof. We consider casework on $t \in [-1,1]$. Recall that

$$\mathcal{I}_{mid} := \{t \mid w(t) = v(t)\} = \left[\sqrt{1 - \tfrac{1}{\pi^2(d+1)^2C^2}}, \sqrt{1 + \tfrac{1}{\pi^2(d+1)^2C^2}}\right]$$

and $\mathcal{I}_{cap} := [-1,1] \setminus \mathcal{I}_{mid}$. By Lemma 7.8, there exists a constant $C_0 \geq 1$ such that $\frac{1}{C_0} \leq \frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} \leq C_0$ for all $t \in \mathcal{I}_{mid}$. By Lemma 7.9 and Lemma 7.10, there exists a constant C_3 such that

$$\frac{C_3}{\log^4 d} \leq \frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} \leq C_3$$

for all $t \in \mathcal{I}_{cap}$. Thus by setting $C_1 = \min\left(C_3, \frac{1}{C_0}\right)$ and $C_2 = \max(C_0, C_3)$, we have that

$$\frac{C_1}{\log^4 d} \leq \frac{\tau[\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}](t)}{w(t)} \leq C_2$$

for all $t \in [-1,1]$. □

Acknowledgments. Cameron Musco was supported by NSF grants 2046235 and 1763618, and an Adobe Research grant. Christopher Musco and Raphael Meyer were supported by NSF grant 2045590 and DOE Award DE-SC0022266. David P. Woodruff and Samson Zhou were supported by a Simons Investigator Award and by the National Science Foundation under Grant No. CCF-1815840. We thank Apoorv Singh and Axel Elaldi for help designing our figures.

References

[AKM+19] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. A universal sampling method for reconstructing signals with simple fourier transforms. In *Proceedings of the Fifty-first Annual ACM on Symposium on Theory of Computing, STOC (2019)*, 2019.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

[BBS94] C. Sidney Burrus, J. A. Barreto, and Ivan W. Selesnick. Iterative reweighted least-squares design of FIR filters. *IEEE Trans. Signal Process.*, 42(11):2926–2936, 1994.

[BDM+20] Vladimir Braverman, Petros Drineas, Cameron Musco, Christopher Musco, Jalaj Upadhyay, David P. Woodruff, and Samson Zhou. Near optimal linear algebra in the online and sliding window models. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 517–528, 2020.

[BLM89] Jean Bourgain, Joram Lindenstrauss, and Vitali Milman. Approximation of zonoids by zonotopes. *Acta mathematica*, 162:73–141, 1989.

[BX09] John P. Boyd and Fei Xu. Divergence (runge phenomenon) for least-squares polynomial approximation on an equispaced grid and mock–chebyshev subset interpolation. *Applied Mathematics and Computation*, 210(1):158–168, 2009.

[CD21] Xue Chen and Michal Derezinski. Query complexity of least absolute deviation regression via robust uniform convergence. In *Conference on Learning Theory, COLT*, pages 1144–1179, 2021.

[CDL13] Albert Cohen, Mark A. Davenport, and Dany Leviatan. On the stability and accuracy of least squares approximations. *Foundations of Computational Mathematics*, 13(5):819–834, 2013.

[CDW18] Graham Cormode, Charlie Dickens, and David P. Woodruff. Leveraging well-conditioned bases: Streaming and distributed summaries in minkowski p-norms. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, pages 1048–1056, 2018.

[CKPS16] Xue Chen, Daniel M. Kane, Eric Price, and Zhao Song. Fourier-sparse interpolation without a frequency gap. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 741–750, 2016.

[CL06] Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, 3(1):79–127, 2006.

[CM17] Albert Cohen and Giovanni Migliorati. Optimal weighted least-squares methods. *The SMAI journal of computational mathematics*, 3:181–203, 2017.

[CP15] Michael B. Cohen and Richard Peng. l_p row sampling by lewis weights. In Rocco A. Servedio and Ronitt Rubinfeld, editors, *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC*, pages 183–192. ACM, 2015.

[CP19a] Xue Chen and Eric Price. Active regression via linear-sample sparsification active regression via linear-sample sparsification. In *Proceedings of the Thirty-Second Conference on Learning Theory (COLT 2019)*, 2019.

[CP19b] Xue Chen and Eric Price. Estimating the frequency of a clustered signal. In *Proceedings of the 46th International Colloquium on Automata, Languages and Programming (ICALP)*, 2019.

[CWW19] Kenneth L. Clarkson, Ruosong Wang, and David P. Woodruff. Dimensionality reduction for tukey regression. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, pages 1262–1271, 2019.

[DDH+08] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for l_p regression. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 932–941, 2008.

[Dum07] Bogdan Dumitrescu. *Positive trigonometric polynomials and signal processing applications*, volume 103. Springer, 2007.

[ELMM20] Yonina C Eldar, Jerry Li, Cameron Musco, and Christopher Musco. Sample efficient toeplitz covariance estimation. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 378–397. SIAM, 2020.

[EMM20] Tamás Erdélyi, Cameron Musco, and Christopher Musco. Fourier sparse leverage scores and approximate kernel learning. *34th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[EN92] Tamás Erdélyi and Paul Nevai. Generalized jacobi weights, christoffel functions, and zeros of orthogonal polynomials. *Journal of approximation theory*, 69(2):111–132, 1992.

[GM99] Narendra K Govil and Ram N Mohapatra. Markov and bernstein type inequalities for polynomials. *Journal of Inequalities and Applications*, 3(4):349–387, 1999.

[HD15] Jerrad Hampton and Alireza Doostan. Coherence motivated sampling and convergence analysis of least squares polynomial chaos regression. *Computer Methods in Applied Mechanics and Engineering*, 290:73–97, 2015.

[KKMS08] Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM J. Comput.*, 37(6):1777–1805, 2008.

[KKP17] Daniel Kane, Sushrut Karmalkar, and Eric Price. Robust polynomial regression up to the information theoretic limit. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 391–402, 2017.

[Lor83] Lee Lorch. Alternative proof of a sharpened form of bernstein's inequality for legendre polynomials. *Applicable Analysis*, 14(3):237–240, 1983.

[Mey22] Raphael A. Meyer. Basic properties of leverage scores. https://randnla.github.io/leverage-score-properties/, August 2022. RandNLA Proof Wiki [Online; accessed 3-November-2022].

[MM20] Raphael A. Meyer and Christopher Musco. The statistical cost of robust kernel hyperparameter turning. In

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS, 2020.

[MMM+22] Raphael Meyer, Cameron Musco, Christopher Musco, David P Woodruff, and Samson Zhou. Fast regression for structured inputs. In *International Conference on Learning Representations (ICLR)*, 2022.

[MMWY22] Cameron Musco, Christopher Musco, David P Woodruff, and Taisuke Yasuda. Active sampling for linear regression beyond the ℓ_2 norm. *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS*, 2022.

[NEM94] Paul Nevai, Tamás Erdélyi, and Alphonse P Magnus. Generalized jacobi weights, christoffel functions, and jacobi polynomials. *SIAM Journal on Mathematical Analysis*, 25(2):602–614, 1994.

[Nev86] Paul Nevai. Géza Freud, orthogonal polynomials and Christoffel functions. a case study. *Journal of Approximation Theory*, 48(1):3 – 167, 1986.

[Pow67] Michael JD Powell. On the maximum errors of polynomial approximations defined by interpolation and by least squares criteria. *The Computer Journal*, 9(4):404–407, 1967.

[PPP21] Aditya Parulekar, Advait Parulekar, and Eric Price. L1 regression with lewis weights subsampling. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM*, volume 207, pages 49:1–49:21, 2021.

[Pra87] Vaughan R. Pratt. Direct least-squares fitting of algebraic surfaces. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH*, pages 145–152. ACM, 1987.

[RW12] Holger Rauhut and Rachel Ward. Sparse Legendre expansions via ℓ1-minimization. *Journal of Approximation Theory*, 164(5):517 – 533, 2012.

[Sar06] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[SV14] Sushant Sachdeva and Nisheeth K. Vishnoi. Faster algorithms via approximation theory. *Found. Trends Theor. Comput. Sci.*, 9(2):125–210, 2014.

[SZ01] Gideon Schechtman and Artem Zvavitch. Embedding subspace of l_p into ℓ_p^n, $0 < p < 1$. *Mathematische Nachrichten*, 227:133–142, 2001.

[Tre12] Lloyd N. Trefethen. *Approximation Theory and Approximation Practice*. Society for Industrial and Applied Mathematics, 2012.

A Operator Sensitivity Sampling

In this section, we show Theorem 4.3, which shows that uniform sampling can achieve a constant factor approximation to the L_p polynomial regression problem.

THEOREM 4.3 RESTATED. *Let $p \geq 1$ and suppose $s_1, \ldots, s_{n_0}$ are drawn uniformly from $[-1, 1]$. Let $\boldsymbol{A} \in \mathbb{R}^{n_0 \times (d+1)}$ be the associated Vandermonde matrix, so that $\boldsymbol{A}_{i,j} = s_i^{j-1}$. Let $\mathbf{b} \in \mathbb{R}^{n_0}$ be the evaluations of f, so that $\mathbf{b}_i = f(s_i)$. For $n_0 = O\left(d^5 2^p p^2 \log d\right)$, there exists a universal constant c such that the sketched solution $\hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p$ satisfies*

$$\|\mathcal{P}\hat{\mathbf{x}} - f\|_p \leq c \min_{\mathbf{x} \in \mathbb{R}^{d+1}} \|\mathcal{P}\mathbf{x} - f\|_p$$

with probability at least $\frac{11}{12}$. Further, let $\varepsilon \in (0, 1)$ and suppose $\|f\|_p \leq C \min_{\mathbf{x}} \|\mathcal{P}\mathbf{x} - f\|_p$. If $n_0 = O\left(\frac{1}{\varepsilon^{O(p^2)}} d^5 p^{O(p)} \log \frac{d}{\varepsilon}\right)$, then

$$\|\mathcal{P}\hat{\mathbf{x}} - f\|_p^p \leq (1 + \varepsilon) \min_{\mathbf{x}} \|\mathcal{P}\mathbf{x} - f\|_p^p$$

with probability at least $\frac{11}{12}$.

Throughout this paper, we use two formulations of Bernstein's inequality in the analysis for general p.

IMPORTED THEOREM A.1. (BERNSTEIN'S INEQUALITY, THEOREMS 3.6 AND 3.7 FROM [CL06]) *Let $X_1, \ldots, X_n$ be independent zero-mean random variables with $|X_i| \leq M$ for all i. Then,*

$$\Pr\left[\left|\sum_{i=1}^n X_i\right| \geq \gamma\right] \leq 2 \exp\left(-\frac{\frac{1}{2}\gamma^2}{\sum_{i=1}^n \mathbb{E}[X_i^2] + \frac{1}{3}Mt}\right)$$

IMPORTED THEOREM A.2. (BOUNDED DIFFERENCES CONCENTRATION, THEOREM 17 FROM [CL06]) *Let $X_1, \ldots, X_n$ be independent random variables such that $|X_i - \mathbb{E}[X_i]| < c_i$ for all $i \in [n]$. Let $X = \sum_i X_i$ and $\gamma > 0$. Then*

$$\Pr[|X - \mathbb{E}[X]| \geq \gamma] \leq \exp\left(\frac{-\gamma^2}{2\sum_i c_i^2}\right)$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

We first show the constant-factor regression guarantee using $O(d^5p^22^p\log(d)$ samples.

LEMMA A.1. *Let $\boldsymbol{A}$ be the Vandermonde matrix formed by sampling $n_0 = O(d^5p^22^p\log(d))$ points from $[-1,1]$ uniformly at random, and let $\mathbf{b}$ be the corresponding observations of f. Then, with probability at least $\frac{11}{12}$, the sketched solution $\hat{\mathbf{x}} := \min_{\mathbf{x}} \|\boldsymbol{A}\mathbf{x}-\mathbf{b}\|_p$ has*

$$\|\mathcal{P}\hat{\mathbf{x}} - f\|_p \le C \min_{\mathbf{x}\in\mathbb{R}^{d+1}} \|\mathcal{P}\mathbf{x} - f\|_p$$

for some universal constant $C > 1$.

Proof. This proof is completed in two standard arguments. First, we show that uniformly sampling enough points yields a ℓ_p subspace embedding, via an $\varepsilon-$Net argument. Second, we use a standard argument that triangle inequality and subspace embedding suffice for constant factor regression [ELMM20, MM20, PPP21].

Let $s_1, ..., s_{n_0}$ denotes the uniformly sampled times. First, fix any vector $\mathbf{x} \in \mathbb{R}^{d+1}$. Then let $Y_i := \frac{2}{n_0}|\mathcal{P}\mathbf{x}(s_i)|^p$, so that $\mathbb{E}[Y_i] = \frac{1}{n_0}\|\mathcal{P}\mathbf{x}\|_p^p$. Note that $|\mathcal{P}\mathbf{x}(s_i)|^p \le d^2(p+1)$ by Lemma 4.2, so that $Y_i \le \frac{2d^2(p+1)}{n_0}\|\mathcal{P}\mathbf{x}\|_p^p$ and therefore $|Y_i - \mathbb{E}[Y_i]| \le \frac{3d^2(p+1)}{n_0}\|\mathcal{P}\mathbf{x}\|_p^p$. Then by letting $Y = \sum_{i=1}^{n_0} Y_i = \|r\boldsymbol{A}\mathbf{x}\|_p^p$, where $r = (\frac{1}{n_0})^{1/p}$ is a rescaling factor, and applying the Bounded Differences Inequality (Imported Theorem A.2) for $\gamma = 2^{-p}\|\mathcal{P}\mathbf{x}\|_p^p$, yields

$$\begin{aligned}
\mathbf{Pr}\left[\left|\|r\boldsymbol{A}\mathbf{x}\|_p^p - \|\mathcal{P}\mathbf{x}\|_p^p\right| \ge 2^{-p}\|\mathcal{P}\mathbf{x}\|_p^p\right] &\le 2\exp\left(\frac{-2^{-p}\|\mathcal{P}\mathbf{x}\|_p^{2p}}{2n_0\frac{9d^4(p+1)^2}{n_0^2}\|\mathcal{P}\mathbf{x}\|_p^{2p}}\right) \\
&= 2\exp\left(\frac{-n_0}{9\cdot 2^p d^4(p+1)^2}\right) \\
&\le \frac{1}{\exp(O(d\log d))}
\end{aligned}$$

Where the last line uses the fact that $n_0 = O(d^5p^22^p\log(d))$. Note that $|a-b| \le |a^p - b^p|^{1/p}$ for all $a, b > 0$ and $p \ge 1$. So, we get $|\|r\boldsymbol{A}\mathbf{x}\|_p - \|\mathcal{P}\mathbf{x}\|_p| \le \left|\|r\boldsymbol{A}\mathbf{x}\|_p^p - \|\mathcal{P}\mathbf{x}\|_p^p\right|^{1/p} \le \frac{1}{2}\|\mathcal{P}\mathbf{x}\|_p$ with high probability. That is,

$$\mathbf{Pr}\left[|\|r\boldsymbol{A}\mathbf{x}\|_p - \|\mathcal{P}\mathbf{x}\|_p| \ge \frac{1}{2}\|\mathcal{P}\mathbf{x}\|_p\right] \le \frac{1}{\exp(O(d\log d))} \tag{A.1}$$

We now union bound this guarantee over a net. We first define the ball $\mathcal{B} = \{\mathbf{x} \mid \|\mathcal{P}\mathbf{x}\|_p = 1\}$. The let $\mathcal{N}$ denote a net over $\mathcal{B}$ such that, for any $\mathbf{x} \in \mathcal{B}$, there exists some $\mathbf{y} \in \mathcal{N}$ such that $\|\mathcal{P}\mathbf{x} - \mathcal{P}\mathbf{y}\|_p \le 0.1$. By Lemma 2.4 of [BLM89], $\mathcal{N}$ has at most $10^{O(d)}$ elements.

Next, note that any $\mathbf{x} \in \mathcal{B}$ can be written as $\mathbf{x} = \sum_{i=0}^{\infty}\alpha_i\mathbf{y}_i$ where $\alpha_0 = 1$ and $|\alpha_i| \le 0.1^i$ and $\mathbf{y}_i \in \mathcal{N}$. So, we can union bound Equation A.1 over all $\mathbf{y} \in \mathcal{N}$ to upper bound

$$\begin{aligned}
\|r\boldsymbol{A}\mathbf{x}\|_p &\le \sum_{i=0}^{\infty}\alpha_i\|r\boldsymbol{A}\mathbf{y}_i\|_p \\
&\le 1.5\sum_{i=0}^{\infty}\alpha_i\|\mathcal{P}\mathbf{y}_i\|_p \\
&\le 1.5\sum_{i=0}^{\infty}0.1^i \\
&= \frac{1.5}{1-0.1} \\
&\le 1.825
\end{aligned}$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

And lower bound

$$\begin{aligned}\|r\boldsymbol{A}\mathbf{x}\|_p &\geq \alpha_0\|r\boldsymbol{A}\mathbf{y}_0\|_p - \sum_{i=1}^{\infty}\alpha_i\|r\boldsymbol{A}\mathbf{y}_i\|_p \\ &\geq (1-0.5)\alpha_0\|\mathcal{P}\mathbf{y}_0\|_p - (1+0.5)\sum_{i=1}^{\infty}\alpha_i\|\mathcal{P}\mathbf{y}_i\|_p \\ &\geq 0.5 - 1.5\sum_{i=1}^{\infty}0.1^i \\ &= 0.3\end{aligned}$$

That is, $\|r\boldsymbol{A}\mathbf{x}\|_p = 1 \pm 0.9$ for any $\mathbf{x}$ such that $\|\mathcal{P}\mathbf{x}\|_p = 1$. So, just by scaling this guarantee, we have shown that for all $\mathbf{x} \in \mathbb{R}^{d+1}$ we have

$$\left|\|r\boldsymbol{A}\mathbf{x}\|_p - \|\mathcal{P}\mathbf{x}\|_p\right| \leq 0.9\|\mathcal{P}\mathbf{x}\|_p$$

This is the complete subspace guarantee. We now bound the error of the sketched solution $\hat{\mathbf{x}}$.

Let $\mathbf{x}^* := \operatorname{argmin}_{\mathbf{x}} \|\mathcal{P}\mathbf{x} - f\|_p$ attain the best optimal loss. Then, by repeated use of the triangle inequality and our subspace embedding,

$$\begin{aligned}\|\mathcal{P}\hat{\mathbf{x}} - f\|_p &\leq \|\mathcal{P}\hat{\mathbf{x}} - \mathcal{P}\mathbf{x}^*\|_p + \|\mathcal{P}\mathbf{x}^* - f\|_p \\ &\leq 2r\|\boldsymbol{A}\hat{\mathbf{x}} - \boldsymbol{A}\mathbf{x}^*\|_p + \|\mathcal{P}\mathbf{x}^* - f\|_p \\ &\leq 2r(\|\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{b}\|_p + \|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p) + \|\mathcal{P}\mathbf{x}^* - f\|_p \\ &\leq 4r\|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p + \|\mathcal{P}\mathbf{x}^* - f\|_p \qquad \text{(Optimality of } \tilde{\mathbf{x}}\text{)}\end{aligned}$$

Then, noting that $\mathbb{E}[\|r(\boldsymbol{A}\mathbf{x}^* - \mathbf{b})\|_p^p] = \|\mathcal{P}\mathbf{x}^* - f\|_p^p$ so that by Markov's inequality we have $\|r(\boldsymbol{A}\mathbf{x}^* - \mathbf{b})\|_p \leq 10\|\mathcal{P}\mathbf{x}^* - f\|_p^p$ with probability 0.9, we conclude that

$$\|\mathcal{P}\hat{\mathbf{x}} - f\|_p \leq 41\|\mathcal{P}\mathbf{x}^* - f\|_p$$

which completes the proof. ∎

We next show that any near-optimal solution to the L_p matrix regression problem formed from subsampling a large number of points in $[-1, 1]$ also corresponds to a near-optimal solution to the L_p polynomial regression problem.

LEMMA A.2. *Let $\boldsymbol{A}$ be the Vandermonde matrix formed by sampling $n_0 = O(d^5p^2 2^p \log(d))$ points on $[-1, 1]$, and let $\mathbf{b}$ be the corresponding observations of f. Let $OPT = \min_{\mathbf{x}\in\mathbb{R}^{d+1}} \|\mathcal{P}\mathbf{x} - f\|_p$. Then with probability at least 0.9, all $\hat{\mathbf{x}} \in \mathbb{R}^{d+1}$ with $\|\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{b}\|_p \leq 11OPT$ have $\|\mathcal{P}\hat{\mathbf{x}} - f\|_p \leq 24OPT$.*

Proof. Let $\mathbf{x}^*$ be a minimizer of $\|\mathcal{P}\mathbf{x} - f\|_p$ so that $OPT = \|\mathcal{P}\mathbf{x}^* - f\|_p$. We now suppose by contradiction that $\|\mathcal{P}\hat{\mathbf{x}} - f\|_p > 24OPT$. By triangle inequality,

$$\|\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{b}\|_p \geq \|\boldsymbol{A}(\hat{\mathbf{x}} - \mathbf{x}^*)\|_p - \|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p.$$

Since $\boldsymbol{A}$ is formed by uniform sampling with $n_0 = \text{poly}(dp/\varepsilon)$ points from $[-1, 1]$, then with high probability,

$$\frac{23}{24}\|\boldsymbol{A}\mathbf{x}\|_p \leq \|\mathcal{P}\mathbf{x}\|_p \leq \frac{25}{24}\|\boldsymbol{A}\mathbf{x}\|_p$$

for all $\mathbf{x} \in \mathbb{R}^{d+1}$. Formally, we prove such a bound in the proof of Lemma A.1. Moreover, note that since $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^{d+1}} \|\mathcal{P}\mathbf{x} - f\|_p$ has $OPT = \|\mathcal{P}\mathbf{x}^* - f\|_p$, then we have $\mathbb{E}[\|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p^p] = OPT^p$. Thus by Jensen's inequality for $p \geq 1$, we have $\mathbb{E}[\|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p] \leq OPT$ and by Markov's inequality,

$$\mathbf{Pr}\left[\|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p \geq 11OPT\right] \leq \frac{1}{11}.$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Thus with probability at least 0.9,

$$\|\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{b}\|_p \geq \frac{23}{24}\|\mathcal{P}(\hat{\mathbf{x}} - \mathbf{x}^*)\|_p - 11OPT.$$

By triangle inequality,

$$\|\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{b}\|_p \geq \frac{23}{24}\|\mathcal{P}\hat{\mathbf{x}} - f\|_p - \|\mathcal{P}\mathbf{x}^* - f\|_p - 11OPT.$$

Thus if $\|\mathcal{P}\hat{\mathbf{x}} - f\|_p > 24OPT$, then

$$\|\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{b}\|_p > 23OPT - OPT - 11OPT = 11OPT,$$

which contradicts the given fact that $\|\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{b}\|_p \leq 11OPT$. Hence we must have

$$\|\mathcal{P}\hat{\mathbf{x}} - f\|_p \leq 24OPT.$$

□

LEMMA A.3. *Let $OPT = \min_{\mathbf{x}\in\mathbb{R}^{d+1}} \|\mathcal{P}\mathbf{x} - f\|_p$ and suppose that $\|f\|_p \leq C \cdot OPT$ for some fixed constant $C \geq 1$. Let $\boldsymbol{A}$ be the Vandermonde matrix formed by sampling $n_0 = O\left(\frac{1}{\varepsilon^{O(p^2)}} d^5 p^{O(p^2)} \log\frac{d}{\varepsilon}\right)$ random points uniformly from $[-1, 1]$. Let $\mathbf{b}$ be the corresponding evaluations of f. Then with probability at least 0.9, the minimizer $\hat{\mathbf{x}}$ to $\min_{\mathbf{x}\in\mathbb{R}^{d+1}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p$ satisfies*

$$\|\mathcal{P}\hat{\mathbf{x}} - f\|_p \leq (1 + \varepsilon)OPT.$$

Proof. We first note that by Lemma 4.2, all sensitivities of $\mathcal{P}$ are at most $M := d^2(p + 1)$.

Note that if $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^{d+1}} \|\mathcal{P}\mathbf{x} - f\|_p$, so that $OPT = \|\mathcal{P}\mathbf{x}^* - f\|_p$, then we have $\mathbb{E}[\|\boldsymbol{A}\mathbf{x}^* - f\|_p^p] = OPT^p$. By Jensen's inequality for $p \geq 1$, we have $\mathbb{E}[\|\boldsymbol{A}\mathbf{x}^* - f\|_p] \leq OPT$. Thus by Markov's inequality,

$$\mathbf{Pr}\left[\|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p \geq 11OPT\right] \leq \frac{1}{11}.$$

We condition against this event. Then, we have $\|\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{b}\|_p \leq \|\boldsymbol{A}\mathbf{x}^* - \mathbf{b}\|_p \leq 11OPT$, so by Lemma A.2 we also have $\|\mathcal{P}\hat{\mathbf{x}} - f\|_p \leq 24OPT$.

For the rest of this proof, let $\mathbf{z} \in \mathbb{R}^{d+1}$ be any vector such that $\|\mathcal{P}\mathbf{z} - f\|_p \leq 24OPT$. By triangle inequality, we have

$$\|\mathcal{P}\mathbf{z}\|_p \leq \|\mathcal{P}\mathbf{z} - f\|_p + \|f\|_p \leq 25C \cdot OPT$$

Since the sensitivities of $\mathcal{P}$ are all at most M, then by definition of sensitivities, we have that for all $u \in [-1, 1]$,

$$\frac{|[\mathcal{P}\mathbf{z}](u)|^p}{\|\mathcal{P}\mathbf{z}\|_p^p} \leq M$$

In particular, for all $u \in [-1, 1]$,

$$|[\mathcal{P}\mathbf{z}](u)|^p \leq \tau := M25^p C^p OPT^p.$$

We partition the points of interval $[-1, 1]$ into two groups. We define $G = \{t : |f(t)|^p \leq \tau p^{2p}/\varepsilon^{p^2}\}$ and $B = \{t : |f(t)|^p > \tau p^{2p}/\varepsilon^{p^2}\}$. Intuitively, B is a set of "bad times" where f is so large that polynomials cannot fit it, and G is the remaining set of "good times". So for any $\mathbf{z}$ with $\|\mathcal{P}\mathbf{z} - f\|_p \leq 24OPT$, we have $|[\mathcal{P}\mathbf{z}](t)| \leq \tau^{1/p}$ as before, and also for any $u \in B$ we have $|f(u)| > \frac{p^2}{\varepsilon^p}\tau^{1/p}$. Thus, for any $u \in B$, we have

$$\left(1 - \frac{\varepsilon^p}{p^2}\right)|f(u)| \leq |[\mathcal{P}\mathbf{z}](u) - f(u)| \leq \left(1 + \frac{\varepsilon^p}{p^2}\right)|f(u)|$$

Therefore,

$$(1 - \varepsilon)|f(u)|^p \leq |[\mathcal{P}\mathbf{z}](u) - f(u)|^p \leq (1 + \varepsilon)|f(u)|^p. \tag{A.2}$$

This formalizes the idea that f cannot be fit by a polynomial on B. On the other hand, for G, we have

$$\|\boldsymbol{A}\mathbf{z} - \mathbf{b}\|_p^p = \|\boldsymbol{A}_G\mathbf{z} - \mathbf{b}_G\|_p^p + \|\boldsymbol{A}_B\mathbf{z} - \mathbf{b}_B\|_p^p, \tag{A.3}$$

where $\boldsymbol{A}_G$ and $\mathbf{b}_G$ are the rows of $\boldsymbol{A}$ and $\mathbf{b}$ associated with points sampled in G, and where $\boldsymbol{A}_B$ and $\mathbf{b}_B$ are similarly the rows associated with points sampled in B. We will next show via an ε-Net argument that the residual $\|\boldsymbol{A}_G\mathbf{z} - \mathbf{b}_G\|_p$ is preserved for all valid $\mathbf{z}$ vectors.

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Accuracy of A on a single coefficient vector z at points in G. For each sample s_i with $i \in [n_0]$, if $s_i \in G$, we define $X_i = \frac{2}{n_0}|\mathcal{P}\mathbf{z}(s_i) - f(s_i)|^p$ to be the corresponding contribution to the empirical residue by the sample. Otherwise, if $s_i \notin G$, we define $X_i = 0$. Since we sample uniformly, i.e., the probability density function for s_i satisfies $p(t) = \frac{1}{2}$ for all $t \in [-1, 1]$, then

$$\mathbb{E}[X_i] = \frac{1}{n_0}\int_{t\in G} |\mathcal{P}\mathbf{z}(t) - f(t)|^p \, dt = \frac{1}{n_0}\|\mathcal{P}\mathbf{z} - f\|_G^p,$$

where $\|f\|_G^p := \int_{t\in G} |f|^p \, dt$ is the integral only over the set G. Because $|\mathcal{P}\mathbf{z}(u)|^p \le \tau$ and $|f(u)|^p \le \tau p^{2p}/\varepsilon^{p^2}$ for all $u \in G$, we have

$$|\mathcal{P}\mathbf{z}(u) - f(u)| \le \frac{2p^2}{\varepsilon^p}\tau^{1/p}$$

so that

$$|X_i - \mathbb{E}[X_i]| \le \frac{2}{n_0}|\mathcal{P}\mathbf{z}(s_i) - f(s_i)|^p + \frac{1}{n_0}\|\mathcal{P}\mathbf{z} - f\|_G^p \le \frac{4}{n_0}\left(\frac{2p^2}{\varepsilon^p}\right)^p \tau$$

Then let $X = \sum_{i\in[n_0]} X_i$ so that, letting $r := (\frac{2}{n_0})^{1/p}$ be a rescaling factor,

$$\mathbb{E}[\|r(\boldsymbol{A}_G\mathbf{z} - \mathbf{b}_G)\|_p^p] = \mathbb{E}[X] = \|\mathcal{P}\mathbf{z} - f\|_G^p$$

Setting $\gamma = \frac{\varepsilon^p}{2^{O(p^2)}}\|f\|_p^p$ in the formulation of Bernstein's concentration inequality in Imported Theorem A.2, we have

$$\mathbf{Pr}\left[|X - \mathbb{E}[X]| \ge \gamma\right] \le \exp\left(\frac{-\frac{\varepsilon^p}{2^{O(p^2)}}\|f\|_p^{2p}}{2\sum_{i\in[n_0]}(\frac{4}{n_0}(\frac{2p^2}{\varepsilon^p})^p\tau)^2}\right) \le \exp\left(\frac{-\varepsilon^{2+2p^2}}{32(50Cp^2)^{2p}M^2 2^{O(p^2)}} \cdot n_0\right).$$

Thus for $n_0 = O\left(\frac{1}{\varepsilon^{2+2p^2}} \, d \cdot M^2 p^{O(p^2)} \log\frac{d}{\varepsilon}\right)$, we have

$$\mathbf{Pr}\left[\left|\|r(\boldsymbol{A}_G\mathbf{z} - \mathbf{b}_G)\|_p^p - \|\mathcal{P}\mathbf{z} - f\|_G^p\right| \ge \frac{\varepsilon^p}{2^{O(p^2)}}\|f\|_p^p\right] \le \frac{1}{\exp(O(dp\log d/\varepsilon))},$$

which implies by concavity (and therefore subadditivity) of $t \mapsto t^{1/p}$ for $p \ge 1$,

$$\mathbf{Pr}\left[\left|\|r(\boldsymbol{A}_G\mathbf{z} - \mathbf{b}_G)\|_p - \|\mathcal{P}\mathbf{z} - f\|_G\right| \ge \frac{\varepsilon}{2^{O(p)}}\|f\|_p\right] \le \frac{1}{\exp(O(dp\log d/\varepsilon))}. \tag{A.4}$$

By a similar argument, let $Y_i := \frac{2}{n_0}|\mathcal{P}\mathbf{z}(s_i)|^p$ for all s_i, so that $\mathbb{E}[Y_i] = \frac{1}{n_0}\|\mathcal{P}\mathbf{z}\|_p^p$ and $|Y_i - \mathbb{E}[Y_i]| \le \frac{3M}{n_0}\|\mathcal{P}\mathbf{z}\|_p^p$. Then $Y := \sum_{i\in[n_0]} Y_i$, by Imported Theorem A.2 for $\gamma = \varepsilon^{p+1}\|\mathcal{P}\mathbf{z}\|_p^p$, yields

$$\mathbf{Pr}\left[\left|\|r\boldsymbol{A}\mathbf{z}\|_p - \|\mathcal{P}\mathbf{z}\|_p\right| \le \varepsilon\|\mathcal{P}\mathbf{z}\|_p\right] \le \frac{1}{\exp(O(d\log d/\varepsilon))}. \tag{A.5}$$

Since $M = (p+1)(d+1)^2$, the total number of samples is $n_0 = O\left(\frac{1}{\varepsilon^{2+2p^2}} \, d^5 p^{O(p^2)} \log\frac{d}{\varepsilon}\right)$.

The arguments so far, when combined carefully (see the last part of this proof), imply that the error from uniform sampling does not matter on B, and that for any fixed $\mathbf{z}$ such that $\|\mathcal{P}\mathbf{z} - f\|_p \le 24OPT$, the error on G is preserved. So, for any such $\mathbf{z}$, we can say with high probability that

$$(1-\varepsilon)\|\mathcal{P}\mathbf{z} - f\|_p^p \le \|\boldsymbol{A}\mathbf{z} - \mathbf{b}\|_p^p \le (1+\varepsilon)\|\mathcal{P}\mathbf{z} - f\|_p^p$$

However, the epsilon-net argument needs to be applied to just G on its own, so we now construct a net under the $\|\cdot\|_G$ norm.

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

ε-net argument for subspace embedding. We now union bound over a net by first defining the ball $\mathcal{B} = \{\mathbf{x} \mid \|\mathcal{P}\mathbf{x}\|_p = 1\}$. The let $\mathcal{N}$ denote a net over $\mathcal{B}$ such that, for any $\mathbf{x} \in \mathcal{B}$, there exists some $\mathbf{y} \in \mathcal{N}$ such that $\|\mathcal{P}\mathbf{x} - \mathcal{P}\mathbf{y}\|_p \leq \varepsilon$. By Lemma 2.4 of [BLM89], $\mathcal{N}$ has at most $(\frac{1}{\varepsilon})^{O(d)}$ elements.

Next, note that any $\mathbf{x} \in \mathcal{B}$ can be written as $\mathbf{x} = \sum_{i=0}^{\infty} \alpha_i \mathbf{y}_i$ where $\alpha_0 = 1$ and $|\alpha_i| \leq \varepsilon^i$ and $\mathbf{y}_i \in \mathcal{N}$. So, we can union bound Equation A.5 over all $\mathbf{y} \in \mathcal{N}$ to upper bound

$$\begin{aligned}
\|r\boldsymbol{A}\mathbf{x}\|_p &\leq \sum_{i=0}^{\infty} \alpha_i \|r\boldsymbol{A}\mathbf{y}_i\|_p \\
&\leq (1+\varepsilon)\sum_{i=0}^{\infty} \alpha_i \|\mathcal{P}\mathbf{y}_i\|_p \\
&\leq (1+\varepsilon)\sum_{i=0}^{\infty} \varepsilon^i \\
&= \frac{1+\varepsilon}{1-\varepsilon} \\
&\leq 1 + 4\varepsilon
\end{aligned}$$

And lower bound

$$\begin{aligned}
\|r\boldsymbol{A}\mathbf{x}\|_p &\geq \alpha_0 \|r\boldsymbol{A}\mathbf{y}_0\|_p - \sum_{i=1}^{\infty} \alpha_i \|r\boldsymbol{A}\mathbf{y}_i\|_p \\
&\geq (1-\varepsilon)\alpha_0 \|\mathcal{P}\mathbf{y}_0\|_p - (1+\varepsilon)\sum_{i=1}^{\infty} \alpha_i \|\mathcal{P}\mathbf{y}_i\|_p \\
&\geq (1-\varepsilon) - (1+\varepsilon)\sum_{i=1}^{\infty} \varepsilon^i \\
&\geq 1 - 6\varepsilon
\end{aligned}$$

That is, $\|r\boldsymbol{A}\mathbf{x}\|_p = 1 \pm 6\varepsilon$ for any $\mathbf{x}$ such that $\|\mathcal{P}\mathbf{x}\|_p = 1$. So, just by scaling this guarantee, we have shown that for all $\mathbf{x} \in \mathbb{R}^{d+1}$ we have

$$\left| \|r\boldsymbol{A}\mathbf{x}\|_p - \|\mathcal{P}\mathbf{x}\|_p \right| \leq 6\varepsilon \|\mathcal{P}\mathbf{x}\|_p$$

ε-net argument over all coefficient vectors. Now again consider any $\mathbf{z}$ such that $\|\mathcal{P}\mathbf{z} - f\|_p \leq 24OPT$. Then $\|\mathcal{P}\mathbf{z}\|_p \leq 25\|f\|_p$:

$$\|\mathcal{P}\mathbf{z}\|_p \leq \|\mathcal{P}\mathbf{z} - f\|_p + \|f\|_p \leq 24 \min_{\mathbf{x}} \|\mathcal{P}\mathbf{x} - f\|_p + \|f\|_p \leq 25\|f\|_p$$

Then let $\mathbf{y} = \alpha\mathbf{y}_0$ be the corresponding net vector as in the previous paragraph. Then we have $\|\mathcal{P}\mathbf{y} - f\|_p \leq 26OPT$ for $\varepsilon \leq O(1)$:

$$\|\mathcal{P}\mathbf{y} - f\|_p \leq \|\mathcal{P}\mathbf{z} - f\|_p + \|\mathcal{P}\mathbf{z} - \mathcal{P}\mathbf{y}\|_p \leq 24OPT + 6\varepsilon\|\mathcal{P}\mathbf{x}'\|_p \leq 24OPT + 6\varepsilon 21\|f\|_p \leq (20 + 21 \cdot 6\varepsilon C)OPT$$

Let $\mathcal{N}'$ be a net over $\mathcal{B}' := \{\mathbf{z} \mid \|\mathcal{P}\mathbf{z}\|_p \leq 26OPT\}$ such that any $\mathbf{z} \in \mathcal{B}'$ has some $\mathbf{y} \in \mathcal{B}'$ such that $\|\mathcal{P}\mathbf{z} - \mathcal{P}\mathbf{y}\|_p \leq \frac{\varepsilon}{2^{O(p)}} \cdot 26OPT$. Since $\frac{1}{26OPT}\mathcal{N}'$ is a ε-Net for the unit ball in the range of $\mathcal{P}$, Lemma 2.4 from [BLM89] tells us that this net has size $(\frac{2^{O(p)}}{\varepsilon})^{O(d)}$. We union bound Equation A.4 over all $\mathbf{y} \in \mathcal{N}'$.

Then, we can write $\mathbf{z} = \sum_{i=0}^{\infty} \alpha_i \mathbf{y}_i$ with $\alpha_0 = 1$, $\alpha_i \leq \varepsilon$, and $\mathbf{y}_i \in \mathcal{N}'$. We will then write $\mathbf{z} = \mathbf{y}_0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} := \sum_{i=1}^{\infty} \alpha_i \mathbf{y}_i$ and apply the triangle inequality:

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

$$\begin{aligned}
\left|\|r(\boldsymbol{A}_G\mathbf{z}-\mathbf{b}_G)\|_p - \|\mathcal{P}\mathbf{z}-f\|_G\right| &\le \left|\|r(\boldsymbol{A}_G\mathbf{y}_0-\mathbf{b}_G)\|_p - \|\mathcal{P}\mathbf{y}_0-f\|_G\right| + \|r\boldsymbol{A}_G\boldsymbol{\Delta}_0\|_p + \|\mathcal{P}\boldsymbol{\Delta}_0\|_G \\
&\le \left|\|r(\boldsymbol{A}_G\mathbf{y}_0-\mathbf{b}_G)\|_p - \|\mathcal{P}\mathbf{y}_0-f\|_G\right| + \|r\boldsymbol{A}\boldsymbol{\Delta}_0\|_p + \|\mathcal{P}\boldsymbol{\Delta}_0\|_p \\
&\le \left|\|r(\boldsymbol{A}_G\mathbf{y}_0-\mathbf{b}_G)\|_p - \|\mathcal{P}\mathbf{y}_0-f\|_G\right| + (1+6\varepsilon)\|\mathcal{P}\boldsymbol{\Delta}_0\|_p \\
&\le \tfrac{\varepsilon}{2^{O(p)}}\|f\|_p + (1+6\varepsilon)\sum_{i=1}^{\infty}(\tfrac{\varepsilon}{2^{O(p)}})^i \cdot 26OPT \\
&\le O(\tfrac{\varepsilon}{2^{O(p)}})\|f\|_p
\end{aligned}$$

In other words, we have that with high probability for all $\mathbf{z}$ with $\|\mathcal{P}\mathbf{z}-f\|_p \le 24OPT$,

$$\|\mathcal{P}\mathbf{z}-f\|_G - O(\tfrac{\varepsilon}{2^{O(p)}})\|f\|_p \le \|r(\boldsymbol{A}_G\mathbf{z}-\mathbf{b}_G)\|_p \le \|\mathcal{P}\mathbf{z}-f\|_G + O(\tfrac{\varepsilon}{2^{O(p)}})\|f\|_p \tag{A.6}$$

Now we extend Equation A.6 to holds for norms with the exponent of p on them. We do this by case analysis, with either $\|\mathcal{P}\mathbf{z}-f\|_G \le \frac{1}{2}\|f\|_p$ or $\|\mathcal{P}\mathbf{z}-f\|_G \ge \frac{1}{2}\|f\|_p$. If $\|\mathcal{P}\mathbf{z}-f\|_G \le \frac{1}{2}\|f\|_p$, then we use the bound $(u+\varepsilon)^p \le u^p + 2\varepsilon p$ for $u+\varepsilon \le 1$, as proven later in Lemma A.4:

$$\begin{aligned}
\|\boldsymbol{A}_G\mathbf{z}-\mathbf{b}_G\|_p^p &\le (\|\mathcal{P}\mathbf{z}-f\|_G + O(\tfrac{\varepsilon}{2^{O(p)}})\|f\|_p)^p \\
&= \|f\|_p^p(\tfrac{\|\mathcal{P}\mathbf{z}-f\|_G}{\|f\|_p} + O(\tfrac{\varepsilon}{2^{O(p)}}))^p \\
(\tfrac{\|\mathcal{P}\mathbf{z}-f\|_G}{\|f\|_p} + O(\tfrac{\varepsilon}{2^{O(p)}}) \le 1) \quad &\le \|f\|_p^p(\tfrac{\|\mathcal{P}\mathbf{z}-f\|_G^p}{\|f\|_p^p} + O(\tfrac{\varepsilon}{2^{O(p)}}p)) \\
&= \|\mathcal{P}\mathbf{z}-f\|_G^p + O(\tfrac{\varepsilon}{2^{O(p)}}p)\|f\|_p^p \\
&\le \|\mathcal{P}\mathbf{z}-f\|_G^p + O(\varepsilon)\|f\|_p^p
\end{aligned}$$

and similarly the lower bound is

$$\begin{aligned}
\|\boldsymbol{A}_G\mathbf{z}-\mathbf{b}_G\|_p^p &\ge (\|\mathcal{P}\mathbf{z}-f\|_G - O(\tfrac{\varepsilon}{2^{O(p)}})\|f\|_p)^p \\
&= \|f\|_p^p(\tfrac{\|\mathcal{P}\mathbf{z}-f\|_G}{\|f\|_p} - O(\tfrac{\varepsilon}{2^{O(p)}}))^p \\
(\tfrac{\|\mathcal{P}\mathbf{z}-f\|_G}{\|f\|_p} + O(\tfrac{\varepsilon}{2^{O(p)}}) \le 1) \quad &\ge \|f\|_p^p(\tfrac{\|\mathcal{P}\mathbf{z}-f\|_G^p}{\|f\|_p^p} - O(\tfrac{\varepsilon}{2^{O(p)}}p)) \\
&= \|\mathcal{P}\mathbf{z}-f\|_G^p - O(\tfrac{\varepsilon}{2^{O(p)}}p)\|f\|_p^p \\
&\ge \|\mathcal{P}\mathbf{z}-f\|_G^p - O(\varepsilon)\|f\|_p^p
\end{aligned}$$

Which completes the first case. For the second case, where $\|\mathcal{P}\mathbf{z}-f\|_G \ge \frac{1}{2}\|f\|_p$ so that $\frac{\|f\|_p}{\|\mathcal{P}\mathbf{z}-f\|_G} \le 2$, we use the bound $(1\pm u)^p \in 1 \pm p(2e)^{p/2}u$ for $u \in [0,1]$, as proven later in Lemma A.4:

$$\begin{aligned}
\|\boldsymbol{A}_G\mathbf{z}-\mathbf{b}_G\|_p^p &\in (\|\mathcal{P}\mathbf{z}-f\|_G \pm O(\tfrac{\varepsilon}{2^{O(p)}})\|f\|_p)^p \\
&= \|\mathcal{P}\mathbf{z}-f\|_G^p(1 + O(\tfrac{\varepsilon}{2^{O(p)}})\tfrac{\|f\|_p}{\|\mathcal{P}\mathbf{z}-f\|_G})^p \\
&\in \|\mathcal{P}\mathbf{z}-f\|_G^p(1 \pm O(\tfrac{\varepsilon}{2^{O(p)}}2^{O(p)})\tfrac{\|f\|_p}{\|\mathcal{P}\mathbf{z}-f\|_G}) \\
&= \|\mathcal{P}\mathbf{z}-f\|_p \pm O(\varepsilon)\|f\|_p\|\mathcal{P}\mathbf{z}-f\|_G^{p-1} \\
(\|\mathcal{P}\mathbf{z}-f\|_G \le C\|f\|_p) \quad &\in \|\mathcal{P}\mathbf{z}-f\|_p \pm O(\varepsilon)\|f\|_p^p
\end{aligned}$$

Which concludes the case analysis, and we find that all $\mathbf{z}$ with $\|\mathcal{P}\mathbf{z}-f\|_p \le 24OPT$ have

$$\|\mathcal{P}\mathbf{z}-f\|_G^p - O(\varepsilon)\|f\|_p^p \le \|r(\boldsymbol{A}_G\mathbf{z}-\mathbf{b}_G)\|_p^p \le \|\mathcal{P}\mathbf{z}-f\|_G^p + O(\varepsilon)\|f\|_p^p$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Finishing the argument. Recall that the interval $[-1, 1]$ is partitioned into two groups G and B and that we analyze the samples s_i with $i \in [n_0]$ depending on whether $s_i \in G$ or $s_i \in B$. Moreover, recall that by Equation A.3, we have

$$\|\boldsymbol{A}\mathbf{z} - \mathbf{b}\|_p^p = \|\boldsymbol{A}_G\mathbf{z} - \mathbf{b}_G\|_p^p + \|\boldsymbol{A}_B\mathbf{z} - \mathbf{b}_B\|_p^p$$

where $\boldsymbol{A}_G$ and $\mathbf{b}_G$ contain the points in G while $\boldsymbol{A}_B$ and $\mathbf{b}_B$ contain the points in B. For any $\mathbf{z}$ with $\|\mathcal{P}\mathbf{z} - f\|_p \leq 24OPT$ and $u \in B$, we have by Equation A.2,

$$(1-\varepsilon)|f(u)|^p \leq |\mathcal{P}\mathbf{z}(u) - f(u)|^p \leq (1+\varepsilon)|f(u)|^p.$$

Since this loss is independent of the value of $\mathbf{z}$, we can view $\sum_{i:s_i\in B}|f(s_i)|^p$ effectively as the sample error of any $\mathbf{z}$ on the bad set. Since $\mathbb{E}[\sum_{i:s_i\in B}|f(s_i)|^p] = \sum_{i=1}^{n_0}\mathbb{E}[\mathbb{1}_{[s_i\in B]}|f(s_i)|^p] = \frac{n_0}{2}\|f\|_B^p$, we get $\sum_{i:s_i\in B}|f(s_i)|^p \leq 50n_0\|f\|_B^p$ with probability $\frac{99}{100}$ by Markov's Inequality. Recalling that $r^p = \frac{1}{n_0}$, we get

$$\begin{aligned}
\|r(\boldsymbol{A}_B\mathbf{z} - \mathbf{b}_B)\|_p^p &= \sum_{i:s_i\in B}\frac{1}{n_0}|\mathcal{P}\mathbf{z}(s_i) - f(s_i)|^p \\
&\in \sum_{i:s_i\in B}\frac{1}{n_0}\left(|f(s_i)|^p \pm \varepsilon|f(s_i)|^p\right) \\
&= \sum_{i:s_i\in B}\frac{1}{n_0}|f(s_i)|^p \pm \varepsilon\left(\frac{1}{n_0}\sum_{i:s_i\in B}|f(s_i)|^p\right) \\
&\subseteq \|r\mathbf{b}_B\|_p^p \pm O(\varepsilon)\|f\|_B^p \\
&\subseteq \|r\mathbf{b}_B\|_p^p \pm O(\varepsilon)\|f\|_p^p
\end{aligned}$$

Next recall that for any $\mathbf{z}$ with $\|\mathcal{P}\mathbf{z} - f\|_p \leq 24OPT$, we have

$$\begin{aligned}
\|\boldsymbol{A}_G\mathbf{z} - \mathbf{b}_G\|_p^p &\geq \|\mathcal{P}\mathbf{z} - f\|_G^p - O(\varepsilon)\|f\|_p^p \\
\|\boldsymbol{A}_G\mathbf{z} - \mathbf{b}_G\|_p^p &\leq \|\mathcal{P}\mathbf{z} - f\|_G^p + O(\varepsilon)\|f\|_p^p
\end{aligned}$$

Thus, the minimizer $\hat{\mathbf{x}}$ to $\min_{\mathbf{x}\in\mathbb{R}^{d+1}}\|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_p$ and $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^{d+1}}\|\mathcal{P}\mathbf{x} - f\|_p$ must satisfy

$$\begin{aligned}
\|\mathcal{P}\hat{\mathbf{x}} - f\|_G^p &\leq \|r(\boldsymbol{A}_G\hat{\mathbf{x}} - \mathbf{b}_G)\|_p^p + O(\varepsilon)\|f\|_p^p \\
&= \|r(\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{b})\|_p^p - \|r(\boldsymbol{A}_B\hat{\mathbf{x}} - \mathbf{b}_B)\|_p^p + O(\varepsilon)\|f\|_p^p \\
&\leq \|r(\boldsymbol{A}\hat{\mathbf{x}} - \mathbf{b})\|_p^p - \|r\mathbf{b}_B\|_p^p + O(\varepsilon)\|f\|_p^p \\
&\leq \|r(\boldsymbol{A}\mathbf{x}^* - \mathbf{b})\|_p^p - \|r\mathbf{b}_B\|_p^p + O(\varepsilon)\|f\|_p^p \\
&\leq \|r(\boldsymbol{A}\mathbf{x}^* - \mathbf{b})\|_p^p - \|r(\boldsymbol{A}_B\mathbf{x}^* - \mathbf{b}_B)\|_p^p + O(\varepsilon)\|f\|_p^p \\
&= \|r(\boldsymbol{A}_G\mathbf{x}^* - \mathbf{b}_G)\|_p^p + O(\varepsilon)\|f\|_p^p \\
&\leq \|\mathcal{P}\mathbf{x}^* - f\|_G^p + O(\varepsilon)\|f\|_p^p
\end{aligned}$$

Since $\|f\|_p = O(OPT)$, it follows that

$$\|\mathcal{P}\hat{\mathbf{x}} - f\|_G^p \leq \|\mathcal{P}\mathbf{x}^* - f\|_G^p + O(\varepsilon C^p)OPT^p$$

Finally, since for $u \in B$ we have both

$$\begin{aligned}
(1-\varepsilon)|f(u)|^p \leq |\mathcal{P}\hat{\mathbf{x}}(u) - f(u)|^p \leq (1+\varepsilon)|f(u)|^p \\
(1-\varepsilon)|f(u)|^p \leq |\mathcal{P}\mathbf{x}^*(u) - f(u)|^p \leq (1+\varepsilon)|f(u)|^p
\end{aligned}$$

by Equation A.2, it then follows that

$$\int_{t\in B}|\mathcal{P}\hat{\mathbf{x}}(t) - f(t)|^p\,dt \leq \int_{t\in B}|\mathcal{P}\mathbf{x}^*(t) - f(t)|^p\,dt + O(\varepsilon C^p)OPT^p$$

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Therefore, we have

$$\int_{-1}^{1} |\mathcal{P}\hat{\mathbf{x}}(t) - f(t)|^p \, dt \leq O(\varepsilon C^p) OPT^p$$

The claim then follows from rescaling ε to $\frac{\varepsilon}{C^p}$. ∎

LEMMA A.4. *Fix $u \geq 0$, $\varepsilon \geq 0$, and even integer $p \geq 1$. If $u + \varepsilon \leq 1$, then $(u+\varepsilon)^p \leq u^p + 2\varepsilon p$. If $u \in [0,1]$, then $(1 \pm u) \in 1 \pm p(2e)^{p/2} u$.*

Proof. Since $u + \varepsilon \leq 1$ and $u^p + 2\varepsilon p \geq 1$ for $p \geq \frac{1}{\varepsilon}$, then we have that

$$(u+\varepsilon)^p \leq 1 \leq u^p + 2\varepsilon p,$$

for all $p \geq \frac{1}{\varepsilon}$ and thus it remains to consider the case where $p < \frac{1}{\varepsilon}$.

To that end, note that by the binomial expansion, we have

$$\begin{aligned}
(u+\varepsilon)^p &= u^p \left(1 + \frac{\varepsilon}{u}\right)^p \\
&= u^p \left(1 + \binom{p}{1}\frac{\varepsilon}{u} + \binom{p}{2}\left(\frac{\varepsilon}{u}\right)^2 + \binom{p}{3}\left(\frac{\varepsilon}{u}\right)^3 + \ldots + \frac{\varepsilon^p}{u^p}\right) \\
&\leq u^p \left(1 + \frac{\varepsilon p}{u} + \frac{\varepsilon^2 p^2}{2! u^2} + \frac{\varepsilon^3 p^3}{3! u^3} + \ldots + \frac{\varepsilon^p p^p}{p! u^p}\right).
\end{aligned}$$

For $p < \frac{1}{\varepsilon}$, we thus have

$$\begin{aligned}
(u+\varepsilon)^p &\leq u^p \left(1 + \frac{\varepsilon p}{u} + \frac{\varepsilon^2 p^2}{2! u^2} + \frac{\varepsilon^3 p^3}{3! u^3} + \ldots + \frac{\varepsilon^p p^p}{p! u^p}\right) \\
&< u^p \left(1 + \frac{\varepsilon p}{u} + \frac{\varepsilon p}{2! u^2} + \frac{\varepsilon p}{3! u^3} + \ldots + \frac{\varepsilon p}{p! u^p}\right) \\
&< u^p + u^p \left(\frac{\varepsilon p}{u^p} + \frac{\varepsilon p}{2! u^p} + \frac{\varepsilon p}{3! u^p} + \ldots + \frac{\varepsilon p}{p! u^p}\right) \\
&\leq u^p + \varepsilon p \left(1 + \frac{1}{2!} + \frac{1}{3!} + \ldots + \frac{1}{p!}\right) \\
&< u^p + \varepsilon p \left(1 + \frac{1}{2!} + \frac{1}{3!} + \ldots\right) = u^p + \varepsilon p (e-1) < u^p + 2\varepsilon p,
\end{aligned}$$

as desired.

For the second claim, consider any $u \in [0,1]$.

$$\begin{aligned}
(1+u)^p &= 1 + \sum_{k=1}^{p} \binom{p}{k} u^k \\
\left(\tbinom{p}{k} \leq \tbinom{p}{p/2}\right) \qquad &\leq 1 + \sum_{k-1}^{p} \binom{p}{p/2} u^k \\
\left(\tbinom{p}{k} \leq (\tfrac{pe}{k})^k\right) \qquad &\leq 1 + p(\tfrac{pe}{p/2})^{p/2} u \\
&= 1 + p(pe)^{p/2} u \\
(1-u)^p &= 1 + \sum_{k=1}^{p} \binom{p}{k} (-u)^k \\
&\geq 1 - \sum_{k=1}^{p} \binom{p}{p/2} u^k \\
&\geq 1 - p(pe)p/2u
\end{aligned}$$

∎

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

A.1 Tighter Bounds for L_p Sensitivities

In this section, we show tighter bounds for the L_p sensitivities. While we do not use this result in the paper, we find that it may be useful for future research on polynomial regression. We first require the following results from polynomial approximation theory.

When t is not near the boundaries of the interval $[-1,1]$, we have a sharper upper bound on the magnitude of the derivative when compared to the Markov brothers' inequality.

THEOREM A.1. (BERNSTEIN'S INEQUALITY, E.G., THEOREM 2.8 IN [GM99]) *Suppose $q(t)$ is a polynomial of degree at most d such that $|q(t)| \le 1$ for $t \in [-1,1]$. Then for all $t \in [-1,1]$, $|q'(t)| \le \frac{d}{\sqrt{1-t^2}}$.*

THEOREM A.2. (POLYNOMIAL APPROXIMATION OF THE INVERSE EXPONENTIAL FUNCTION, E.G., THEOREM 4.1 IN [SV14]) *For every $c > 0$ and $\varepsilon \in (0,1]$, there exists a polynomial $q_{c,\varepsilon}$ with degree $O\left(\sqrt{\max\left(c, \log\frac{1}{\varepsilon}\right) \cdot \log\frac{1}{\varepsilon}}\right)$ such that*

$$\max_{x\in[0,c]} |e^{-x} - q_{c,\varepsilon}(x)| \le \varepsilon$$

COROLLARY A.1. (POLYNOMIAL APPROXIMATION OF THE GAUSSIAN KERNEL) *There exists a polynomial q with degree $O(d\log(pd))$ such that $|q(x)| \le 1$ for all $x \in [-2,2]$ and*

$$\max_{x\in[-2,2]} |e^{-20d^2\log(d)x^2} - q(x)| \le \frac{1}{pd^4}$$

Proof. By Theorem A.2, there exists a polynomial $\hat{q}$ of degree $O(d\log(pd))$ such that

$$\max_{x\in[0,80d^2\log(d)]} \left|e^{-x} - \hat{q}(x)\right| \le \frac{1}{2pd^4}$$

By taking the polynomial $\tilde{q}(x) = \hat{q}(20d^2\log(d)x^2)$, we find a polynomial q with degree $O(d\log d)$ such that

$$\max_{x\in[-2,2]} \left|e^{-20d^2\log(d)x^2} - q(x)\right| \le \frac{1}{2pd^4}$$

Then, since $0 \le e^{-4d^2\log(d)x^2} \le 1$, it suffices to take $q(t) = \tilde{q}(t) - \frac{1}{2pd^4}$ to ensure $|q(x)| \le 1$ for all $x \in [-2,2]$. ∎

We now prove an upper bound on the L_p sensitivities that is linear in d in the interior of the interval $[-1,1]$, which crucially improves upon known quadratic bounds, e.g., Lemma 4.2.

THEOREM A.3. (UPPER BOUND ON SENSITIVITY) *Let $p \ge 1$ be a fixed constant and let q be a polynomial of degree at most $d \ge 12$. Then for $t \in [-1,1]$, the L_p sensitivity of t satisfies*

$$\psi_p[\mathcal{P}](t) := \max_{\deg(q)\le d} \frac{|q(t)|^p}{\int_{-1}^{1} |q(x)|^p\, dx} \le d^2(p+1).$$

Moreover for $|t| \le 1 - \frac{1}{d}$, the L_p sensitivity of t satisfies

$$\psi_p[\mathcal{P}](t) := \max_{\deg(q)\le d} \frac{|q(t)|^p}{\int_{-1}^{1} |q(x)|^p\, dx} = O\left(\frac{dp\log(dp)}{\sqrt{1-t^2}}\right)$$

Proof. The first bound is just Lemma 4.2 restated, which directly relied on the Markov brothers' bound. So here we show that for $|t| \le 1 - \frac{1}{d}$, we have

$$\frac{|q(t)|^p}{\int_{-1}^{1} |q(x)|^p\, dx} = O\left(\frac{dp\log(dp)}{\sqrt{1-t^2}}\right)$$

For this sharper bound on the sensitivity in the middle of the interval $[-1,1]$, we need a more sophisticated argument.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Let $q := \operatorname{argmax}_{\|q\|_\infty=1} \frac{|q(t)|^p}{\|q\|_p^p}$ maximize the sensitivity at t.

We would ideally like to use Bernstein's Inequality (Theorem A.1) to lower bound the mass of q around t, much like the proof of Lemma 4.2. Indeed, if q were maximized at t, then such a proof would be as simple as Lemma 4.2. That proof picks any t^* such that $|q(t^*)| = 1$ and lower bounds $\|q\|_p^p$ by integrating over an interval of length $\frac{1}{d^2}$ around t^*. Crucially, this is tight because even if t^* is far from t, the Markov Brothers' bound on $|q'(x)|$ is independent of x. However, Bernstein's bound $|q'(x)| \le \frac{d}{\sqrt{1-x^2}}$ would give very different results depending on how far t^* is from t. So, the weight of this proof is in showing that q must be maximized near t.

We first show that $q(t)$ is not terribly small. Since $\|q\|_\infty = 1$, by the proof of Lemma 4.2, we know that $\|q\|_p^p \ge \frac{1}{d^2(p+1)}$. Let $c(t) := 1$ be the constant function. Since q maximizes the sensitivity function, we get have $\frac{|q(t)|^p}{\|q\|_p^p} \ge \frac{|c(t)|^p}{\|c\|_p^p} = \frac{1}{2}$. So, $|q(t)|^p \ge \frac{\|q\|_p^p}{2} \ge \frac{1}{2d^2(p+1)}$. Without loss of generality $q(t) > 0$, so we just write $q(t) \ge \frac{1}{d^{2/p}(2p+2)^{1/p}} \ge \frac{1}{4d^2}$ (since $(2p+2)^{1/p} \le 4$ for $p \ge 1$).

Next, we multiply q with a degree $\tilde{O}(d)$ polynomial approximation of a Gaussian pdf centered at t, which effectively erases q outside of a small interval of t. Intuitively, this negligibly changes the degree of q but ensures that the maximum is achieved near t.

We first argue that multiplying by an exact Gaussian bump maximizes q near t. Let $a(x) := e^{-4(x-t)^2 d^2 \log(d)}$ be this Gaussian bump. Let $C := q(t)$. By Markov Brothers', $|q'(x)| \le d^2$. So, we can bound the growth of q around t:

$$|q(t+x)| \le C + d^2 |x| \le 1 + d^2 |x|$$

Scaling by the Gaussian,

$$|a(t+x)q(t+x)| \le (1 + d^2 |x|)e^{-5x^2 d^2 \log(d)}$$

For $|x| > \frac{1}{2d}$, we have $e^{-20x^2 d^2 \log(d)} < e^{-5 \log d} = \frac{1}{d^5}$. So, for $|x| > \frac{1}{2d}$,

$$|a(t+x)q(t+x)| \le \frac{1}{d^5}(1 + d^2 |x|) \le \frac{3}{d^3}$$

And since $a(t)q(t) = q(t) = C > \frac{1}{4d^2}$, we guarantee that $\operatorname{argmax}_{[-1,1]} a(x)q(x) \in [t - \frac{1}{2d}, t + \frac{1}{2d}]$ for $d \ge 12$.

Next, we substitute $a(x)$ with the polynomial approximation $b(x)$ guaranteed by Corollary A.1. Namely, we know that $b(x)$ has degree at most $\xi d \log(pd)$ for some constant $\xi > 1$, and that $|b(x) - a(x)| \le \frac{1}{d^4}$ for all $x \in [-1, 1]$. Then, we get that $f(x) := b(x)q(x)$ is a degree $d + \xi d \log(pd) \le 2\xi d \log(pd)$ polynomial with $f(t) \ge \frac{1}{4d^2} - \frac{1}{d^4}$ and $|f(t+x)| \le \frac{3}{d^3} + \frac{1}{d^4}$, so that f is still maximized in $[t - \frac{1}{2d}, t + \frac{1}{2d}]$.

Now, we can appeal to Bernstein's bound to control the sensitivity of f. Let $r(x) := \frac{1}{f(t)} f(x)$ be a rescaling of f so that $\|r\|_\infty = \frac{1}{f(t)}$ and $r(t) = 1$. By Bernstein's bound, we have $|r'(x)| \le \frac{2\xi d \log(pd)}{f(t)\sqrt{1-x^2}}$, and therefore that $|r'(t+x)| \le \frac{4\xi d \log(pd)}{f(t)\sqrt{1-t^2}}$ for $x \in [0, \frac{1}{2d}]$ (via the smoothness of the Chebyshev measure – see Lemma E.3 in the appendix). Let $m_t := \frac{4\xi d \log(pd)}{\sqrt{1-t^2}}$ be this locally accurate bound on the derivative (but without $f(t)$), and let $t^* := \operatorname{argmax}_{[-1,1]} r(x) \in [t - \frac{1}{2d}, t + \frac{1}{2d}]$, so that we have:

$$r(t^* + x) \ge \frac{1}{f(t)} - \frac{m_t}{f(t)} x \ge 1 - m_t x \ge 0 \qquad \forall x \in [0, \tfrac{1}{m_t}]$$

which follows since $f(t) \le q(t) \le 1$. Then, we get

$$\|r\|_p^p \ge \int_0^{1/m_t} (1 - m_t x)^p dx = \frac{1}{m_t(p+1)}$$

So that

$$\frac{|f(t)|^p}{\|f\|_p^p} = \frac{|r(t)|^p}{\|r\|_p^p} \le \frac{1}{\frac{1}{m_t(p+1)}} = \frac{4\xi(p+1)d \log(pd)}{\sqrt{1-t^2}}$$

Then, since $|b(x)| \le 1$, we have $|q(x)| \ge |f(x)|$, so that $\|q\|_p^p \ge \|f\|_p^p$. Further, since $q(t) = \frac{f(t)}{b(t)} \le \frac{f(t)}{1 - \frac{1}{pd^4}}$, we get $|q(t)|^p \le (1 - \frac{1}{pd^4})^{-p} |f(t)|^p \le 2 |f(t)|^p$. We conclude:

$$\frac{|q(t)|^p}{\|q\|_p^p} \le 2 \frac{|f(t)|^p}{\|f\|_p^p} \le \frac{8\xi(p+1)d \log(pd)}{\sqrt{1-t^2}}$$

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

□

We also offer the following lower bound on the L_p sensitivities.

LEMMA A.5. (LOWER BOUND ON SENSITIVITY) *For any $t \in [-1,1]$, $p \geq 1$, and $d \geq \Omega(p)$ we have*

$$\psi_p[\mathcal{P}](t) = \max_{q:\deg(q)\leq d} \frac{|q(t)|^p}{\int_{-1}^{1} |q(x)|^p \, dx} = \Omega\left(\frac{d\sqrt{p}}{\sqrt{\log d}}\right)$$

Proof. Let $t \in [-1,1]$. By Corollary A.1, there exists a polynomial q with degree d such that

$$\max_{x\in[-1,1]} |e^{-(Cd/\sqrt{\log d})^2(x-t)^2} - q(x)| \leq \frac{1}{d},$$

for a fixed constant $C > 0$. Let $f(x) = e^{-(Cd/\sqrt{\log d})^2(x-t)^2}$ so that $f(t) = 1$ and for $p \geq 1$,

$$\int_{-1}^{1} |f(x)|^p \, dx = \int_{-1}^{1} (f(x))^p \, dx < \int_{-\infty}^{\infty} e^{-p(Cd/\sqrt{\log d})^2 x^2} \, dx = \frac{\sqrt{\pi \log d}}{Cd\sqrt{p}}$$

Hence, $\frac{|f(t)|^p}{\int_{-1}^{1} |f(x)|^p \, dx} \geq \frac{Cd\sqrt{p}}{\sqrt{\pi \log d}}$. Since $\max_{x\in[-1,1]} |e^{-(Cd/\sqrt{\log d})^2(x-t)^2} - q(x)| \leq \frac{1}{d}$, then it follows that

$$|q(t)| \geq \frac{1}{2}$$

and

$$\|q - f\|_p^p \leq \int_{-1}^{1} \frac{1}{d^p} dt = \frac{2}{d^p}$$

Therefore,

$$\begin{aligned}
\|q\|_p^p &\leq (\|f\|_p + \|q - f\|_p)^p \\
&\leq \left(\left(\frac{\sqrt{\pi \log d}}{Cd\sqrt{p}}\right)^{1/p} + \frac{2^{1/p}}{d}\right)^p \\
&\leq \left(\left(1 + \frac{1}{p^{\frac{p-1}{p}}}\right)\left(\frac{\sqrt{\pi \log d}}{Cd\sqrt{p}}\right)^{1/p}\right)^p \\
&= \left(1 + \frac{1}{p^{\frac{p-1}{p}}}\right)^p \cdot \frac{\sqrt{\pi \log d}}{Cd\sqrt{p}} \\
&\leq 2\frac{\sqrt{\pi \log d}}{Cd\sqrt{p}}
\end{aligned}$$

where the third inequality comes showing that $\frac{2^{1/p}}{d} \leq \frac{1}{p^{\frac{p-1}{p}}}(\frac{\sqrt{\pi \log d}}{Cd\sqrt{p}})^{1/p}$, which holds when $d \geq p(2C\sqrt{p})^{\frac{1}{p-1}} = \Omega(p)$. We therefore conclude that

$$\frac{|q(t)|^p}{\int_{-1}^{1} |q(x)|^p \, dx} \geq \frac{\frac{1}{2}}{2\frac{\sqrt{\pi \log d}}{Cd\sqrt{p}}} = \frac{d\sqrt{p}}{4\sqrt{\pi \log(d)}}$$

□

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

B Reweighted Operator L_2 Subspace Embedding

THEOREM B.1. *Suppose $s_1, \dots, s_{n_0}$ are drawn uniformly from $[-1,1]$. Let $\boldsymbol{A} \in \mathbb{R}^{n_0 \times (d+1)}$ be the associated Vandermonde matrix, so that $\boldsymbol{A}_{i,j} = s_i^{j-1}$. Let $\gamma := \frac{2}{n_0}$. Let $\boldsymbol{W} \in \mathbb{R}^{n_0 \times n_0}$ be diagonal with $\boldsymbol{W}_{ii} = \gamma w(s_i)$. Then for $n_0 = \Omega(d^4 \log d)$, we have that with probability at least $\frac{11}{12}$,*

$$\frac{1}{2}\mathcal{P}^\top \mathcal{W}^{1-\frac{2}{p}} \mathcal{P} \preceq \gamma^{-\frac{2}{p}} \boldsymbol{A}^\top \boldsymbol{W}^{1-\frac{2}{p}} \boldsymbol{A} \preceq 2\mathcal{P}^\top \mathcal{W}^{1-\frac{2}{p}} \mathcal{P},$$

where $\mathcal{W}$ is the operator that rescales by the truncated Chebyshev density $w(t)$ and $p \in [1,2]$.

To prove this claim, it will be more convenient to shift where the γ term is located, into the matrix $\boldsymbol{A}$:

THEOREM B.2. *Suppose $s_1, \dots, s_{n_0}$ are drawn uniformly from $[-1,1]$, and we construct the associated scaled Vandermonde matrix $\boldsymbol{A} \in \mathbb{R}^{n_0 \times (d+1)}$, so that $\boldsymbol{A}_{i,j} = (\frac{2}{n_0})^{\frac{1}{2}} s_i^{j-1}$. Let $\boldsymbol{W} \in \mathbb{R}^{n_0 \times n_0}$ be the diagonal matrix with $\boldsymbol{W}_{i,i} = w(t_i)$. If $n_0 = \Omega(d^5 \log d)$, then with probability at least $\frac{11}{12}$, we have*

$$\frac{1}{2}\mathcal{P}^\top \mathcal{W}^{1-\frac{2}{p}} \mathcal{P} \preceq \boldsymbol{A}^\top \boldsymbol{W}^{1-\frac{2}{p}} \boldsymbol{A} \preceq 2\mathcal{P}^\top \mathcal{W}^{1-\frac{2}{p}} \mathcal{P},$$

where $\mathcal{W}$ is the operator that rescales by the truncated Chebyshev density $w(t)$ and $p \in [\frac{2}{3}, 2]$.

Proof. We first prove a more general statement by considering a general operator $\mathcal{W}$, which we eventually set to be the Lewis weight operator. Let $\mathcal{W} : L_2([-1,1]) \to \mathbb{R}$ be any operator so that $\max_{t \in [-1,1], \mathbf{x} \in \mathbb{R}^{d+1}} \frac{|\mathcal{W}\mathcal{P}\mathbf{x}(t)|^2}{\|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2} \le \mathfrak{S}$ for some $\mathfrak{S} < \infty$. Consider a fixed $\mathbf{x} \in \mathbb{R}^{d+1}$ and suppose we uniformly sample $n_0 = O\left(\frac{\mathfrak{S}^2}{\varepsilon^2} \log \frac{d}{\varepsilon}\right)$ points from $[-1,1]$. For each $i \in [n_0]$, let X_i be the random variable with value $|[\boldsymbol{W}\boldsymbol{A}\mathbf{x}](i)| = \frac{2}{n_0}|\mathcal{W}\mathcal{P}\mathbf{x}(s_i)|^2$. Then $\mathbb{E}[X_i] = \frac{1}{n_0}\|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2$. Moreover, since $|\mathcal{W}\mathcal{P}\mathbf{x}(t)|^2 \le \mathfrak{S} \cdot \|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2$, we get $|X_i - \mathbb{E}[X_i]| \le \frac{1}{n_0}(2\mathfrak{S}+1)\|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2$. Let $X = \sum_{i \in [n_0]} X_i = \|\boldsymbol{W}\boldsymbol{A}\mathbf{x}\|_2^2$ so that by linearity of expectation, $\mathbb{E}[X] = \|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2$. Setting $\gamma = \varepsilon\|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2$ in the formulation of Bernstein's concentration inequality in Imported Theorem A.2, we thus have

$$\mathbf{Pr}\left[|\|\boldsymbol{W}\boldsymbol{A}\mathbf{x}\|_2^2 - \|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2| \le \varepsilon\|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2\right] \le \exp(-O(d\log d/\varepsilon))$$

for any fixed $\mathbf{x} \in \mathbb{R}^{d+1}$. That is, we have

$$(1-\varepsilon)\|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2 \le \frac{1}{n_0}\|\mathbf{W}\mathbf{A}\mathbf{x}\|_2^2 \le (1+\varepsilon)\|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2,$$

with probability at least $1 - \exp(-O(d\log d/\varepsilon))$.

ε-net argument over all coefficient vectors. We now union bound over an ε-net over all coefficient vectors $\mathbf{x} \in \mathbb{R}^{d+1}$. Let $\mathcal{B} = \left\{\mathcal{W}\mathcal{P}\mathbf{x} \;\middle|\; \|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2 \le 1\right\}$. By Lemma 2.4 of [BLM89], we construct a net $\mathcal{N}$ over $\mathcal{B}$ by greedily adding in points that are within L_2 distance $\left(\frac{\varepsilon}{d}\right)$, so that $|\mathcal{N}| \le \left(\frac{d}{\varepsilon}\right)^{O(d)} = e^{O(d\log d/\varepsilon)}$. Note that since we have $(1+\varepsilon)$-accuracy for any $\mathcal{W}\mathcal{P}\mathbf{x} \in \mathcal{N}$ with probability $1 - \exp(-O(d\log d/\varepsilon))$ by sampling $n_0 = O\left(\frac{\mathfrak{S}^2}{\varepsilon^2}\log\frac{d}{\varepsilon}\right)$ points from $[-1,1]$, then by a union bound, we have $(1+\varepsilon)$-accuracy for all points in $\mathcal{N}$ with high probability.

For any $\mathcal{W}\mathcal{P}\mathbf{x}$ with $\|\mathcal{W}\mathcal{P}\mathbf{z}\|_2 = 1$, we construct a sequence $\mathcal{W}\mathcal{P}\mathbf{y}_1, \mathcal{W}\mathcal{P}\mathbf{y}_2, \dots$ such that $\|\mathcal{W}\mathcal{P}\mathbf{z} - \sum_{j=1}^{i}\mathcal{W}\mathcal{P}\mathbf{y}_j\|_2 \le \varepsilon^i$ and such that there exists constants $\gamma_i \le \varepsilon^{i-1}$ with $\frac{1}{\gamma_i}\mathcal{W}\mathcal{P}\mathbf{y}_i \in \mathcal{N}$ for all i. Formally, we let $\mathcal{W}\mathcal{P}\mathbf{y}_1$ be the point in the net $\mathcal{N}$ that is closest to $\mathcal{W}\mathcal{P}\mathbf{x}$, so that $\|\mathcal{W}\mathcal{P}\mathbf{z} - \mathcal{W}\mathcal{P}\mathbf{y}_1\|_2 \le \varepsilon$. We can then define the remaining points $\mathcal{W}\mathcal{P}\mathbf{y}_i$ inductively: For a sequence $\mathcal{W}\mathcal{P}\mathbf{y}_1, \dots, \mathcal{W}\mathcal{P}\mathbf{y}_{i-1}$ such that $\gamma_i := \|\mathcal{W}\mathcal{P}\mathbf{s} - \sum_{j=1}^{i-1}\mathcal{W}\mathcal{P}\mathbf{y}_j\|_2 \le \varepsilon^{i-1}$, observe that $\frac{1}{\gamma_i}\|\mathcal{W}\mathcal{P}\mathbf{s} - \sum_{j=1}^{i-1}\mathcal{W}\mathcal{P}\mathbf{y}_j\|_2 = 1$. Thus, there exists a function $\mathcal{W}\mathcal{P}\mathbf{y}_i \in \mathcal{N}$ that is within distance ε of $\mathcal{W}\mathcal{P}\mathbf{s} - \sum_{j=1}^{i-1}\mathcal{W}\mathcal{P}\mathbf{y}_j$, which completes the induction.

Therefore, for the matrix $\mathbf{A}$ sampled by the algorithm, by triangle inequality,

$$|\|\mathcal{W}\mathcal{P}\mathbf{x}\|_2 - \|\mathbf{W}\mathbf{A}\mathbf{x}\|_2| \le \sum_{i=1}^{\infty}|\|\mathcal{W}\mathcal{P}\mathbf{y}_i\|_2 - \|\mathbf{W}\mathbf{A}\mathbf{y}_i\|_2| \le \sum_{i=1}^{\infty}\varepsilon^i\|\mathcal{W}\mathcal{P}\mathbf{y}_i\|_2 = O(\varepsilon)\|\mathcal{W}\mathcal{P}\mathbf{x}\|_2$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

The correctness over all $\mathbf{x} \in \mathbb{R}^{d+1}$ then follows from a rescaling of ε. Hence, we have that with high probability all $\mathbf{x} \in \mathbb{R}^{d+1}$ enjoy

$$(1-\varepsilon)\|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2 \le \|\mathbf{W}\mathbf{A}\mathbf{x}\|_2^2 \le (1+\varepsilon)\|\mathcal{W}\mathcal{P}\mathbf{x}\|_2^2$$

or equivalently

$$(1-\varepsilon)\mathcal{P}^\top\mathcal{W}\mathcal{P} \preceq \boldsymbol{A}^\top \boldsymbol{W}^{1-\frac{2}{p}}\boldsymbol{A} \preceq (1+\varepsilon)\mathcal{P}^\top\mathcal{W}\mathcal{P}$$

Finishing the argument for the Chebyshev density. Observe that the Chebyshev density satisfies $w(t) \in [d,(d+1)^2]$ for each $t \in [-1,1]$. Since $\max_{t\in[-1,1],q:\deg(q)\le d} \frac{|q(t)|^2}{\|q\|_2^2} \le O(d^2)$, then by substituting the Lewis weight operator $\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}$ in place of the general operator $\mathcal{W}$ in the above analysis, we have that

$$\mathfrak{S} = \max_{t\in[-1,1],\mathbf{x}\in\mathbb{R}^{d+1}} \frac{|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}(t)|^p}{\|\mathcal{W}^{\frac{1}{2}-\frac{1}{p}}\mathcal{P}\mathbf{x}\|_p^p} \le O(d^3) \max_{t\in[-1,1],\mathbf{x}\in\mathbb{R}^{d+1}} \frac{|\mathcal{P}\mathbf{x}(t)|}{\|\mathcal{P}\mathbf{x}\|_2^2} \le O(d^5 p)$$

for the operator $\mathcal{W}$ that corresponds to the Chebyshev weights. Hence the claim follows by taking $\varepsilon = O(1)$. ∎

C From Two-Stage Sampling to One-Stage Sampling

LEMMA C.1. *Fix parameter n_0 and function $f : [-1,1] \to [0,1]$. Suppose $s_1,\dots,s_{n_0}$ are drawn iid uniformly from $[-1,1]$, and we sample biased coins $c_i \sim B(1,f(s_i))$ for $i = 1,\dots,n_0$. Then, the marginal distribution of $\{s_i|c_i = 1\}$ is a distribution with $B(n_0, \frac{1}{2}\int_{-1}^1 f(\tau)d\tau)$ i.i.d. samples with PDF proportional to f.*

Proof. For intuition, we can think of the event $c_i = 1$ as indicating that sample i is accepted. Then $\{s_i|c_i = 1\}$ is the set of time samples returned by this rejection sampling scheme. We now formalize this intuition.

Let p_t denote the PDF of the t variables. We first write simplify two probabilities for a fixed $i \in [n_0]$. First we expand the marginal distribution of the coins:

$$\begin{aligned}\Pr[c_i = 1] &= \int_{-1}^1 \Pr[c_i = 1 \mid s_i = \tau]p_t(\tau)d\tau \\ &= \frac{1}{2}\int_{-1}^1 f(\tau)d\tau \\ &= \frac{1}{2}\|f\|_1\end{aligned}$$

Since each coin is marginally distributed as a $B(1,\frac{1}{2}\|f\|_1)$ random variable, and the number of items in the set $\{t_i|c_i = 1\}$ is the sum of the coins, we conclude that $|\{t_i|c_i = 1\}| \sim B(n_0,\frac{1}{2}\|f\|_1)$.

Let $p_{t_i|c_i=1}$ denote the PDF for t_i when conditioned on $c_i = 1$. Using Bayes' Theorem for continuous and discrete random variables,

$$\begin{aligned}p_{t|c_i=1}(\tau) &= \frac{\Pr[c_i = 1|t_i = \tau]\cdot p_t(\tau)}{\Pr[c_i = 1]} \\ &= \frac{f(\tau)\cdot\frac{1}{2}}{\frac{1}{2}\|f\|_1} \\ &= \frac{f(\tau)}{\int_{-1}^1 f(s)ds}\end{aligned}$$

Thus, each item in $\{t_i|c_i = 1\}$ (which are trivially independent of each-other), is distributed with PDF proportional to f. ∎

LEMMA 4.3 RESTATED. *Suppose n_0 time samples are drawn uniformly from $[-1,1]$, and each sample is thrown away with probability $1-\min\{\frac{m}{n_0}\frac{1}{1-s_i^2},1\}$. Let n denote the number of remaining samples. Then n is distributed as $B(n_0,O(\frac{m}{n_0}))$, and with probability $\frac{99}{100}$ the resulting samples cannot be distinguished from iid samples from the Chebyshev measure.*

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Proof. Below this paragraph, we isolate a simple probability theory claim. Specifically, we apply the below lemma with $f(t) = \min\{1, \frac{m}{n_0}\frac{1}{\sqrt{1-t^2}}\}$, so that the remaining number of samples are distributed as $B(n_0, \frac{m}{2n_0})$. Since the total variation distance between sampling 1 time with respect to f and with respect to v is $O((\frac{m}{n_0})^2)$, the distance between m i.i.d. samples from the two distributions is $O(\frac{m^3}{n_0^2}) = \frac{1}{\tilde{O}(d^7)}$. So, the difference in success probability of our algorithm and one that samples by f is at most $\frac{1}{\tilde{O}(d^7)} \leq \frac{1}{100}$. ∎

D Compact Rounding

LEMMA D.1. *Let $\boldsymbol{B} \in \mathbb{R}^{n_0 \times d_B}$ and $q \in [0,2]$. Let $\mathbf{v} \in \mathbb{R}^{n_0}$ with $|\mathbf{v}(i)| \leq \frac{1}{\varepsilon}(w_q[\boldsymbol{B}](i))^{1/q}$. Let $\mathcal{N}_\varepsilon$ be an ε-Net over $\|\boldsymbol{B}\mathbf{y}\|_q = 1$ with $\log|\mathcal{N}_\varepsilon| = O(d\log(\frac{1}{\varepsilon}))$. Let $\ell = \log_{1+\varepsilon}((2d_B)^{1/q})$. Then, there exists sets of vectors $\mathcal{D}_0, \ldots, \mathcal{D}_\ell \subseteq \mathbb{R}^{n_0}$, such that: For all $\mathbf{u} \in \mathcal{N}_\varepsilon$ we can define $\mathbf{r} := \mathbf{u} - \mathbf{v}$ and pick $\mathbf{d}_0 \in \mathcal{D}_0, \ldots, \mathbf{d}_\ell \in \mathcal{D}_\ell$ to create a "compact rounding" $\mathbf{r}' = \sum_{k=0}^{\ell} \mathbf{d}_k$ where:*

1. $|\mathbf{r}(i) - \mathbf{r}'(i)| \leq \varepsilon \max\{|\mathbf{u}(i)|, |\mathbf{v}(i)|\}$ *for all* $i \in [n_0]$
2. $|\mathbf{d}_k(i)| \leq \frac{2}{\varepsilon}(\frac{1}{2}(\frac{w_q[\boldsymbol{B}](i)}{d_B} + \frac{1}{n_0}))^{1/q}(1+\varepsilon)^{k+2}$ *for all* $i \in [n_0], k \in \{0, \ldots, \ell\}$
3. $\mathbf{d}_0, \ldots, \mathbf{d}_\ell$ *all have disjoints supports*

Further, we have that the sets $\mathcal{D}_0, \ldots, \mathcal{D}_\ell$ are not too large:

$$\log|\mathcal{D}_k| \leq C_r \frac{d_B \log(n_0)}{\varepsilon^{1+q}(1+\varepsilon)^{qk}}$$

To prove Lemma D.1, we need the following structural statement from [MMWY22], attributed to Corollary 4.7 and Proposition in [BLM89] as well as Proposition 3.1 and Remark 3.2 in [SZ01].

LEMMA D.2. (ENTROPY ESTIMATES, [BLM89, MMWY22, SZ01]) *Let $\boldsymbol{B} \in \mathbb{R}^{n_0 \times d_B}$ with $n_0 \geq d_B$ and let $q \in (0,2)$ be a fixed constant. Let $\boldsymbol{W} \in \mathbb{R}^{n_0 \times n_0}$ be the diagonal matrix with $\boldsymbol{W}_{i,i} = \frac{1}{2}\left(\frac{w_q[\boldsymbol{B}](i)}{d_B} + \frac{1}{n_0}\right)$. Let $\mathcal{B}_q = \{\boldsymbol{B}\mathbf{y} : \|\boldsymbol{B}\mathbf{y}\|_q \leq 1\}$. Then for any $\gamma \in [1, d^{1/q}]$, there exists a net $\mathcal{N}_\infty \subset \mathbb{R}^{n_0}$ such that for any $\mathbf{u} \in \mathcal{B}_q$, there exists $\mathbf{f} \in \mathcal{N}_\infty$ with $\|\boldsymbol{W}^{-1/q}(\mathbf{u} - \mathbf{f})\|_\infty \leq \gamma$ and*

$$\log|\mathcal{N}_\infty| \leq c_q \cdot \frac{d_B \log n_0}{\gamma^q}$$

where c_q is a constant that only depends on q.

We next define the index sets and state a structural property on the index sets. The following proof is almost identical to the structural property on the index sets by [MMWY22].

LEMMA D.3. (INDEX SETS, [BLM89, MMWY22]) *For each $k \in \{0, \ldots, \ell\}$, let $\mathcal{N}_k$ be the net defined by Lemma D.2 for $\gamma = \frac{1}{3}\varepsilon(1+\varepsilon)^k > 1$. Otherwise if $\gamma \leq 1$, let $\mathcal{N}_k = \mathcal{N}_\varepsilon$. For each $\mathbf{u} \in \mathcal{N}_\varepsilon$ and $k \in \{0, \ldots, \ell\}$, let $\mathbf{f}_{k,\mathbf{u}} \in \mathcal{N}_k$ satisfy*

$$\|\boldsymbol{W}^{-1/q}(\mathbf{f}_{k,\mathbf{u}} - \mathbf{u})\|_\infty \leq \tfrac{1}{3}\varepsilon(1+\varepsilon)^k$$

as defined in Lemma D.2. Define the index sets

$$C_{k,\mathbf{u}} := \left\{ i \in [n] \;\middle|\; W_{ii}^{-1/q} |\mathbf{f}_{k,\mathbf{u}}(i)| \geq (1+\varepsilon)^{k-1} \right\}$$
$$D_{k,\mathbf{u}} := C_{k,\mathbf{u}} \setminus \bigcup_{k'>k} C_{k',\mathbf{u}}$$
$$D_{0,\mathbf{u}} := [n_0] \setminus \bigcup_{k \geq 1} C_{k,\mathbf{u}}$$

Then for each k, we have $\log|\mathcal{N}_k| = O\left(\frac{d\log n_0}{(\varepsilon(1+\varepsilon)^k)^q}\right)$ and for every $i \in D_{k,\mathbf{u}}$, we have

$$W_{ii}^{1/q}(1+\varepsilon)^{k-2} \leq |\mathbf{u}(i)| \leq W_{ii}^{1/q}(1+\varepsilon)^{k+2}$$

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Proof. First note that the largest γ value we use is $\frac{1}{3}\varepsilon(1+\varepsilon)^{\ell} = \frac{\varepsilon}{3}(2d)^{1/q} \le d^{1/q}$, so we can safely create all of these $\mathcal{N}_k$ sets. Because $\|\boldsymbol{W}^{-1/q}(\mathbf{f}_{k,\mathbf{u}} - \mathbf{u})\|_\infty \le \frac{1}{3}\varepsilon(1+\varepsilon)^k$, we get that all $i \in C_{k,\mathbf{u}}$ have

$$\begin{aligned}|\mathbf{u}(i)| &\ge |\mathbf{f}_{k,\mathbf{u}}(i)| - \tfrac{1}{3}W_{ii}^{1/q}\varepsilon(1+\varepsilon)^k \\ &\ge W_{ii}^{1/q}\left((1+\varepsilon)^{k-1} - \tfrac{1}{3}\varepsilon(1+\varepsilon)^k\right) \\ &\ge W_{ii}^{1/q}(1+\varepsilon)^{k-2}\end{aligned}$$

where the second inequality follows from the definition that $|f_{k,\mathbf{u}}(i)| \ge W_{ii}^{1/q}(1+\varepsilon)^{k-1}$ for $i \in C_{k,\mathbf{u}}$. We similarly have $|\mathbf{u}(i)| \le W_{ii}^{1/q}(1+\varepsilon)^{k+2}$ for $i \notin C_{k,\mathbf{u}}$. Hence for $i \in D_{k,\mathbf{u}} = C_{k,\mathbf{u}} \setminus \bigcup_{k'>k+1} C_{k',\mathbf{u}}$ for $k \in [1,\ell)$, we have

$$W_{ii}^{1/q}(1+\varepsilon)^{(k-2)} \le |\mathbf{u}(i)| \le W_{ii}^{1/q}(1+\varepsilon)^{k+2}$$

as desired. We next show this bound holds for all $i \in D_{\ell,\mathbf{u}}$. The lower bound follows from the same argument as above, but the upper bound needs to be argued separately since there is no $C_{\ell+1,\mathbf{u}}$. We do this by going through the sensitivity bounds on $\mathbf{u}$, which is in the range of $\boldsymbol{B}$:

$$\begin{aligned}|\mathbf{u}(i)| &\le (\psi_q[\boldsymbol{B}](i))^{1/q}\|\mathbf{u}\|_q \\ &\le (w_q[\boldsymbol{B}](i))^{1/q} && \text{(Lemma 3.8 from [CWW19])} \\ &= \left(\frac{w_q[\boldsymbol{B}](i)}{2d_B}\right)^{1/q}(1+\varepsilon)^{\ell} && (\ell = \log_{1+\varepsilon}(2d_B)^{1/q}) \\ &\le W_{ii}^{1/q}(1+\varepsilon)^{\ell+2}\end{aligned}$$

Lemma 3.8 from [CWW19] simply says that for $q \in [1,2]$, the ℓ_q sensitivities lower bound the ℓ_q Lewis weights. This then completes the bulk of the proof since $\frac{1}{2d_B}w_q[\boldsymbol{B}](i) \le \frac{1}{2}(\frac{w_q[\boldsymbol{B}](i)}{d_B} + \frac{1}{n_0}) = W_{ii}$. As a last note, when $\gamma \le 1$, or equivalently $k \le \log_{1+\varepsilon}\frac{3}{\varepsilon}$, we take $f_{k,\mathbf{u}} = \mathbf{u}$ since $\mathcal{N}_k = \mathcal{N}_\varepsilon$, which trivially gives $\|\boldsymbol{W}^{-1/q}(\mathbf{f}_{k,\mathbf{u}} - \mathbf{u})\|_\infty \le \frac{1}{3}\varepsilon(1+\varepsilon)^k$. Further $\log|\mathcal{N}_k| = \log|\mathcal{N}_\varepsilon| \le O(d\log\frac{1}{\varepsilon}) \le O(d\log(n_0)) = O(\frac{d\log(n_0)}{(\varepsilon(1+\varepsilon)^k)^q})$ by Lemma 2.4 of [BLM89], and since $\frac{1}{\varepsilon} \le n_0$. ∎

We similarly define index sets on the measurement vector $\mathbf{v}$, though we remark that the definition is conceptual and not algorithmic, in the sense that the entries of $\mathbf{v}$ do not need to be read.

LEMMA D.4. (INDEX SETS FOR $\mathbf{v}$) *[MMWY22] Fix some $\mathbf{u} \in \mathcal{N}_\varepsilon$. Consider the following sets:*

$$\begin{aligned}&\text{(for } k \in \{0,\dots,\ell\}) && B_{k,\mathbf{u}} := \left\{i \in D_{k,\mathbf{u}} : |\mathbf{v}(i)| \le \tfrac{1}{\varepsilon}W_{ii}^{1/q}(1+\varepsilon)^{k+2}\right\} \\ &\text{(for } k \in \{1,\dots,\ell\}) && H_k := \left\{i \in [n_0] : \tfrac{1}{\varepsilon}W_{ii}^{1/q}(1+\varepsilon)^{k+1} < |\mathbf{v}(i)| \le \tfrac{1}{\varepsilon}W_{ii}^{1/q}(1+\varepsilon)^{k+2}\right\} \\ &\text{(for } k \in \{1,\dots,\ell\}) && G_{k,\mathbf{u}} := H_k \setminus \bigcup_{k' \ge k} C_{k',\mathbf{u}}\end{aligned}$$

Then $B_{0,\mathbf{u}},\dots,B_{\ell,\mathbf{u}},G_{1,\mathbf{u}},\dots,G_{\ell,\mathbf{u}}$ form a partition of $[n_0]$.

Proof. We first prove that $B_{0,\mathbf{u}},\dots,B_{\ell,\mathbf{u}},G_{1,\mathbf{u}},\dots,G_{\ell,\mathbf{u}}$ are disjoint. We then prove that their union is $[n_0]$.

First note that $D_{0,\mathbf{u}},\dots,D_{\ell,\mathbf{u}}$ are clearly disjoint by their definition. Further, since $D_{0,\mathbf{u}}$ is defined by subtracting away all other $D_{k,\mathbf{u}}$ from $[n_0]$, we know that the union of all $D_{0,\mathbf{u}},\dots,D_{\ell,\mathbf{u}}$ is $[n_0]$. That is, $D_{0,\mathbf{u}},\dots,D_{\ell,\mathbf{u}}$ partition $[n_0]$.

Now consider any k,k'. Then,

- $B_{k,\mathbf{u}} \bigcap B_{k',\mathbf{u}} = \emptyset$ since $i \in B_{k,\mathbf{u}}$ implies $i \in D_{k,\mathbf{u}}$ implies $i \notin D_{k',\mathbf{u}}$ implies $i \notin B_{k',\mathbf{u}}$.
- $G_{k,\mathbf{u}} \bigcap G_{k,\mathbf{u}} \subseteq H_k \bigcap H_{k'} = \emptyset$ since H_k and $H_{k'}$ have no intersection by definition.
- For $k \ge k'$, $B_{k,\mathbf{u}} \bigcap G_{k',\mathbf{u}} = \emptyset$ since $i \in B_{k,\mathbf{u}} \subseteq D_{k,\mathbf{u}} \subseteq C_{k,\mathbf{u}}$ implies $i \in \bigcup_{k'' \ge k'} C_{k',\mathbf{u}}$, so that $i \notin G_{k',\mathbf{u}}$ by definition.

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

- For $k < k'$, $B_{k,\mathbf{u}} \bigcap G_{k',\mathbf{u}} = \emptyset$ since $k' \geq k+1$ and $i \in G_{k',\mathbf{u}} \subseteq H_{k'}$ means $|\mathbf{x}(i)| > \frac{1}{\varepsilon} W_{ii}^{1/q}(1+\varepsilon)^{k'+1} \geq \frac{1}{\varepsilon} W_{ii}^{1/q}(1+\varepsilon)^{k+2}$, which contradicts $i \in B_{k,\mathbf{u}} \subseteq D_{k,\mathbf{u}}$.

So, $B_{0,\mathbf{u}}, \ldots, B_{\ell,\mathbf{u}}, G_{1,\mathbf{u}}, \ldots, G_{\ell,\mathbf{u}}$ are disjoint.

Now, consider any $i \in [n_0]$. Then there exists some k such that $i \in D_{k,\mathbf{u}}$. If $|\mathbf{v}(i)| \leq \frac{1}{\varepsilon} W_{ii}^{1/q}(1+\varepsilon)^{k+1}$, then we immediately get that $i \in B_{k,\mathbf{u}}$. Otherwise, if $|\mathbf{v}(i)| > \frac{1}{\varepsilon} W_{ii}^{1/q}(1+\varepsilon)^{k+1}$, then there exists some $k' > k$ such that $i \in H_{k'}$. Notably, this uses the fact that H_ℓ can contain the largest entries of $|\mathbf{v}(i)|$, since $|\mathbf{v}(i)| \leq \frac{1}{\varepsilon}(w_q[\boldsymbol{B}](i))^{1/q} \leq \frac{1}{\varepsilon} W_{ii}^{1/q}(1+\varepsilon)^{\ell+2}$. Since $i \in D_{k,\mathbf{u}} = C_{k,\mathbf{u}} \setminus \bigcup_{k''>k} C_{k'',\mathbf{u}}$, we know that $i \notin C_{k'',\mathbf{u}}$ for any $k'' > k$. Therefore, we know that $i \in G_{k',\mathbf{u}}$, since both $i \in H_{k'}$ and $i \in \bigcup_{k'' \geq k'} C_{k'',\mathbf{u}}$. Therefore, all $i \in [n_0]$ belongs to exactly one of $B_{0,\mathbf{u}}, \ldots, B_{\ell,\mathbf{u}}, G_{1,\mathbf{u}}, \ldots, G_{\ell,\mathbf{u}}$. In other words, $B_{0,\mathbf{u}}, \ldots, B_{\ell,\mathbf{u}}, G_{1,\mathbf{u}}, \ldots, G_{\ell,\mathbf{u}}$ partitions $[n_0]$. ∎

LEMMA D.5. (ℓ_∞ ERROR BOUND) *[MMWY22] Fix some $\mathbf{u} \in \mathcal{N}_\varepsilon$. Then we let $\mathbf{r}' = \mathbf{e} + \sum_{k=0}^{\ell} \mathbf{d}_k$ with $\mathbf{e}$ and $\mathbf{d}_k$ as follows:*

$$\begin{aligned}
\mathbf{d}_0(i) &:= \mathbf{u}(i) - \mathbf{v}(i) && i \in B_{0,\mathbf{u}} \\
\mathbf{d}_k(i) &:= \mathbf{f}_{k,\mathbf{u}}(i) - \mathbf{v}(i) && k \in [\ell], i \in B_{k,\mathbf{u}} \\
\mathbf{d}_k(i) &:= (1+\varepsilon)^k \cdot W_{ii}^{1/q} - \mathbf{v}(i) && k \in [\ell], i \in B_{k,\mathbf{u}} \\
\mathbf{d}_k(i) &:= -\mathbf{v}(i) && i \in G_{k,\mathbf{u}} \\
\mathbf{d}_k(i) &:= 0 && \textit{otherwise}
\end{aligned}$$

Then $|\mathbf{r}(i) - \mathbf{r}'(i)| \leq \varepsilon \max\{|\mathbf{u}(i)|, |\mathbf{v}(i)|\}$ for all $i \in [n_0]$.

Proof. Fix any $i \in [n_0]$. Since $B_{0,\mathbf{u}}, \ldots, B_{\ell,\mathbf{u}}, G_{1,\mathbf{u}}, \ldots, G_{\ell,\mathbf{u}}$ partition $[n_0]$, it suffices to show that if $i \in B_{0,\mathbf{u}}$ or $i \in B_{k,\ell}$ or $i \in G_{k\ell}$ then $|\mathbf{r}(i) - \mathbf{r}'(i)| \leq \varepsilon \max\{|\mathbf{u}(i)|, |\mathbf{v}(i)|\}$. That is, this proof proceeds by case analysis over these three possible cases. First, if $i \in B_{0,\mathbf{u}}$ then $\mathbf{r}'(i) = \mathbf{u}(i) - \mathbf{v}(i)$, so that $|\mathbf{r}(i) - \mathbf{r}'(i)| = 0$. Second, if $i \in B_{k,\mathbf{u}}$ for $k \geq 1$, then $\mathbf{r}'(i) = \mathbf{f}_{k,\mathbf{u}}(i) - \mathbf{v}(i)$, and since $i \in D_{k,\mathbf{u}}$ we get

$$|\mathbf{r}(i) - \mathbf{r}'(i)| = |\mathbf{f}_{k,\mathbf{u}}(i) - \mathbf{u}(i)| \leq \varepsilon \tfrac{(1+\varepsilon)^2}{3} \cdot W_{ii}^{1/q}(1+\varepsilon)^{k-2} \leq \varepsilon \cdot |\mathbf{u}(i)|$$

Third, if $i \in G_{k,\mathbf{u}}$ then $\mathbf{r}'(i) = -\mathbf{v}(i)$, so that $|\mathbf{r}(i) - \mathbf{r}'(i)| = |\mathbf{u}(i)|$. We have $|\mathbf{v}(i)| \geq \frac{1}{\varepsilon} W_{ii}^{1/q}(1+\varepsilon)^{k+1}$, and since $i \in G_{k,\mathbf{u}}$ implies $i \notin C_{k',\mathbf{u}}$ for all $k' \geq k$, we also have $|\mathbf{u}(i)| \leq W_{ii}^{1/q}(1+\varepsilon)^{k+2}$ So,

$$|\mathbf{r}(i) - \mathbf{r}'(i)| = |\mathbf{u}(i)| \leq W_{ii}^{1/q}(1+\varepsilon)^{k+2} \leq \varepsilon |\mathbf{v}(i)|$$

∎

We now prove the compact rounding of Lemma D.1:

LEMMA D.1 RESTATED. *Let $\boldsymbol{B} \in \mathbb{R}^{n_0 \times d_B}$ and $q \in [0,2]$. Let $\mathbf{v} \in \mathbb{R}^{n_0}$ with $|\mathbf{v}(i)| \leq \frac{1}{\varepsilon}(w_q[\boldsymbol{B}](i))^{1/q}$. Let $\mathcal{N}_\varepsilon$ be an ε-Net over $\|\boldsymbol{B}\mathbf{y}\|_q = 1$ with $\log|\mathcal{N}_\varepsilon| = O(d\log(\frac{1}{\varepsilon}))$. Let $\ell = \log_{1+\varepsilon}((2d_B)^{1/q})$. Then, there exists sets of vectors $\mathcal{D}_0, \ldots, \mathcal{D}_\ell \subseteq \mathbb{R}^{n_0}$, such that: For all $\mathbf{u} \in \mathcal{N}_\varepsilon$ we can define $\mathbf{r} := \mathbf{u} - \mathbf{v}$ and pick $\mathbf{d}_0 \in \mathcal{D}_0, \ldots, \mathbf{d}_\ell \in \mathcal{D}_\ell$ to create a "compact rounding" $\mathbf{r}' = \sum_{k=0}^{\ell} \mathbf{d}_k$ where:*

1. *$|\mathbf{r}(i) - \mathbf{r}'(i)| \leq \varepsilon \max\{|\mathbf{u}(i)|, |\mathbf{v}(i)|\}$ for all $i \in [n_0]$*
2. *$|\mathbf{d}_k(i)| \leq \frac{2}{\varepsilon}(\frac{1}{2}(\frac{w_q[\boldsymbol{B}](i)}{d_B} + \frac{1}{n_0}))^{1/q}(1+\varepsilon)^{k+2}$ for all $i \in [n_0], k \in \{0, \ldots, \ell\}$*
3. *$\mathbf{d}_0, \ldots, \mathbf{d}_\ell$ all have disjoints supports*

Further, we have that the sets $\mathcal{D}_0, \ldots, \mathcal{D}_\ell$ are not too large:

$$\log|\mathcal{D}_k| \leq C_r \frac{d_B \log(n_0)}{\varepsilon^{1+q}(1+\varepsilon)^{qk}}$$

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Proof. Fix any $\mathbf{u} \in \mathcal{N}_\varepsilon$. Observe that the first property follows from Lemma D.5. Moreover, note that $\{\mathbf{d}_0, \dots, \mathbf{d}_\ell\}$ are disjoint by Lemma D.4, since the vectors $\mathbf{d}_k$ only have nonzero support on the indices in $B_{k,\mathbf{u}}$ and $G_{k,\mathbf{u}}$. Hence, the third property holds.

Furthermore, note that $|\mathbf{d}_k(i)| \le |\mathbf{v}(i)| + \max\left(|\mathbf{u}(i)|, (1+\varepsilon)^k W_{ii}^{1/q}\right)$ for $i \in B_{k,\mathbf{u}} \cup G_{k,\mathbf{u}}$. In particular, $|\mathbf{v}(i)| \le \frac{1}{\varepsilon} W_{ii}^{1/q}(1+\varepsilon)^{k+2}$. Similarly, we have from Lemma D.3 that $|\mathbf{u}(i)| \le W_{ii}^{1/q}(1+\varepsilon)^{k+1}$. Therefore, $|\mathbf{d}_k(i)| \le \frac{2}{\varepsilon} W_{ii}^{1/q}(1+\varepsilon)^{k+2}$, which gives the second property.

To bound the number of possible vectors $\mathbf{d}_k$, note that $\mathbf{d}_k$ is a deterministic function of $\mathbf{f}_{k,\mathbf{u}}$, $B_{k,\mathbf{u}}$, and $G_{k,\mathbf{u}}$. So, let $B_k := \{B_{k,\mathbf{u}} : \mathbf{u} \in \mathcal{N}_\varepsilon\}$ be the set of all possible "B" index sets generated by the net $\mathcal{N}_\varepsilon$ at layer k, and similarly let $G_k := \{G_{k,\mathbf{u}} : \mathbf{u} \in \mathcal{N}_\varepsilon\}$. Then, looking across all possible fixings of $\mathbf{u} \in \mathcal{N}_\varepsilon$, each $\mathbf{d}_k$ is deterministic in some $\mathbf{f}_{k,\mathbf{u}} \in \mathcal{N}_k$, in some $\mathcal{S}_1 \in B_k$, and in some $\mathcal{S}_2 \in G_k$. So, the number of possible $\mathbf{d}_k$ is at most

$$|\mathcal{D}_k| = |\{\mathbf{d}_k : \mathbf{u} \in \mathcal{N}_\varepsilon\}| \le |\{(\mathbf{f}_{k,\mathbf{u}}, \mathcal{S}_1, \mathcal{S}_2) : \mathbf{f}_{k,\mathbf{u}} \in \mathcal{N}_k,\, S_1 \in B_k,\, S_2 \in G_k\}| = |\mathcal{N}_k| \cdot |B_k| \cdot |G_k|$$

Next, recall that $B_{k,\mathbf{u}} \subseteq D_{k,\mathbf{u}}$, so $|B_{k,\mathbf{u}}| \le |D_{k,\mathbf{u}}|$. Further, recall that $D_{k,\mathbf{u}} = C_{k,\mathbf{u}} \setminus \bigcup_{k'>k} C_{k',\mathbf{u}}$ and that all $C_{k,\mathbf{u}}$ are deterministic in some vector $\mathbf{f}_{k,\mathbf{u}}$ from the net $\mathcal{N}_k$. So, $E_k := \{D_{k,\mathbf{u}} : \mathbf{u} \in \mathcal{N}_\varepsilon\}$ has[11]

$$|B_k| \le |E_k| = |\{D_{k,\mathbf{u}} : \mathbf{u} \in \mathcal{N}_\varepsilon\}| \le |\{(\mathbf{f}_{k,\mathbf{u}}, \dots, \mathbf{f}_{\ell,\mathbf{u}}) : \mathbf{u} \in \mathcal{N}_\varepsilon\}| \le \prod_{k'=k}^{\ell} |\mathcal{N}_{k'}|$$

By Lemma D.2, we have $\log|\mathcal{N}_{k'}| \le c_q \frac{d_B \log n_0}{(3\varepsilon(1+\varepsilon)^{k'})^q}$, so that

$$\log(|B_k|) \le \sum_{k'=k}^{\ell} 3^q c_q \frac{d_B \log n_0}{(\varepsilon(1+\varepsilon)^{k'})^q} \le 2 \cdot 3^q c_q \frac{d_B \log n_0}{\varepsilon^{1+q}(1+\varepsilon)^{qk}}$$

Where the last inequality comes from bounding $\sum_{k'=k}^{\ell} \frac{1}{(1+\varepsilon)^{qk}} \le \sum_{k'=k}^{\infty} \frac{1}{(1+\varepsilon)^{qk}} = \frac{1}{(1+\varepsilon)^{qk}} \cdot \frac{(1+\varepsilon)^q}{(1+\varepsilon)^q - 1} \le \frac{1}{(1+\varepsilon)^{qk}} \cdot \frac{(1+\varepsilon)^2}{1+\varepsilon-1} \le \frac{2}{\varepsilon(1+\varepsilon)^{qk}}$. Similarly, since $G_{k,\mathbf{u}} = H_k \setminus \bigcup_{k' \ge k} C_{k',\mathbf{u}}$ where H_k is a fixed set independent of $\mathbf{u}$, we again get $\log|G| \le \prod_{k'=k}^{\ell} |\mathcal{N}_{k'}|$, and so we conclude:

$$\log|\mathcal{D}_k| \le \log|\mathcal{N}_k| + \log|B_k| + \log|G_k| \le 6 c_q 3^q \frac{d_B \log n_0}{\varepsilon^{1+q}(1+\varepsilon)^{qk}}$$

which completes the bulk of the proof. ∎

E Smaller Relegated Proofs

E.1 Bounding the Generalized Christoffel Function

LEMMA E.1. *Let $f(s)$ be a differentiable concave function on interval $[a-\Delta, a+\Delta]$. Then, $\int_{a-\Delta}^{a+\Delta} f(s)ds \le 2\Delta f(a)$.*

Proof. First recall that concave functions have $f(s) \le f(a) + f'(a)(s-a)$, so we have

$$\int_{a-\Delta}^{a+\Delta} f(s)dx \le \int_{a-\Delta}^{a+\Delta} \left(f(a) - f'(a)(s-a)\right) ds = 2\Delta f(a) + f'(a) \int_{a-\Delta}^{a+\Delta} \left((s-a)\right) ds = 2\Delta f(a) + 0$$

∎

LEMMA E.2. *The generalized Christoffel function $\lambda_d(\alpha, 2, t) := \min_{q:deg(q)\le d} \frac{\int_{-1}^{1} (q(s))^2 \alpha(s) ds}{(q(t))^2}$, where $\alpha(s) := (1-s^2)^{\frac{1}{p}-\frac{1}{2}}$, has $\lambda_d(z, 2, t) \le \frac{C}{d-1}(1-t^2)^{\frac{1}{p}}$ for some universal constant C, for all $t \in \mathcal{I}_{mid}$, for all $p \in [\frac{2}{3}, 2]$.*

[11] We use E_k instead of D_k to avoid confusion with $\mathcal{D}_k$.

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

Proof. By Theorem 2.1 of [EN92], we know that

$$\lambda_d(\alpha, 2, t) \le C^{\Gamma+3}\alpha_M(t)$$

where C is a universal constant, $\Gamma = \frac{2}{p} - 1 \le 1$, and $\alpha_M(t) := \int_{|s-t|\le\Delta_M(t)} \alpha(s)ds$ where $\Delta_M(s) := \max\{\frac{\sqrt{1-s^2}}{M}, \frac{1}{M^2}\}$ and $M = 1 + \frac{2(d-1)}{\Gamma+3} = 1 + \frac{p}{p+1}(d-1) \in [\frac{d-1}{3}, d]$. To bound the integral within $\alpha_M(t)$, we use the above lemma about concave functions. Since $\frac{d^2}{ds^2}\alpha(s) = -2(\frac{1}{p} - \frac{1}{2})(1-t^2)^{\frac{1}{p}-\frac{5}{2}}((\frac{1}{p} - \frac{3}{2})t - 1) \le 0$ for all $t \in [-1, 1]$ for all $p \ge \frac{2}{3}$, we know that α is concave. Then, since $\alpha(s)$ is concave on $[-1, 1]$, we find that for any t such that $|t| + \Delta_M(t) \le 1$ (for which $|t| \le \sqrt{1 - \frac{4}{M^2}}$ suffices), we get

$$\int_{t-\Delta_M(t)}^{t+\Delta_M(t)} \alpha(s)ds \le 2\Delta_M(t)\alpha(t) = 2\frac{\sqrt{1-t^2}}{M}(1-t^2)^{\frac{1}{p}-\frac{1}{2}} \le \frac{2}{M}(1-t^2)^{\frac{1}{p}}$$

Putting this together,

$$\lambda_d(\alpha, 2, t) \le C^{\Gamma+3}\alpha_M(t) \le C^4\frac{2}{\frac{d-1}{3}}(1-t^2)^{\frac{1}{p}} = \frac{3C^4}{d-1}(1-t^2)^{\frac{1}{p}}$$

□

E.2 Smoothness of the Chebyshev Measure

LEMMA E.3. *Let $x \in (-1, 1)$, and let $y \in (-1 + \Delta, 1 - \Delta)$ for $\Delta = \frac{1-x}{2}$. Then,*

$$\frac{1}{\sqrt{1-y^2}} \le \frac{2}{\sqrt{1-x^2}}$$

In particular, if $x = 1 - \frac{1}{d}$, then we get $y \in [1 - \frac{1}{d}, 1 - \frac{1}{2d}]$.

Proof. WLOG, we consider $x > 0$. Since $x \mapsto \frac{1}{\sqrt{1-x^2}}$ is monotonically increasing on $x > 0$, we just need to show that $\frac{1}{\sqrt{1-(x+\Delta)^2}} \le \frac{2}{\sqrt{1-x^2}}$. Rearranging that equation, we get

$$4\Delta^2 + 8x\Delta + (3x^2 - 3) \le 0$$

Plugging in $\Delta = \frac{1-x}{2}$ and simplifying, we see the bound holds for all $x < 1$. □

E.3 Binomial Approximation

LEMMA E.4. *Let $x > 0$, $p > 2$, and $x \le \frac{1}{p}$. Then,*

$$1 - \tfrac{1}{2}px \le (1-x)^p \qquad\qquad (1+x)^p \le 1 + 3px$$

In other words, $(1 \pm x)^p = 1 + \Theta(px)$.

Proof. Let $f(u) := (1+u)^p$, so that the Taylor Approximation $\tilde{f}(u) := f(0) + f'(0)u = 1 + pu$ has the following residual for $u \in [-x, x]$:

$$\left|f(u) - \tilde{f}(u)\right| \le \max_{\xi\in\mathcal{I}} \frac{f''(\xi)}{2}u^2 = \frac{1}{2}p(p-1)u^2 \max_{\xi\in\mathcal{I}}(1+\xi)^{p-2} \le \frac{1}{2}p^2u^2 \max_{\xi\in\mathcal{I}}(1+\xi)^{p-2}$$

Where $\mathcal{I} = [0, x]$ if $u \ge 0$ and where $\mathcal{I} = [-x, 0]$ if $u < 0$. For $u < 0$, we have $\max_{\xi\in[-x,0]}(1+\xi)^{p-2} = 1$, so that $\left|f(u) - \tilde{f}(u)\right| \le \frac{1}{2}p^2u^2$. So, $|(1-x)^p - (1-px)| = \left|f(-x) + \tilde{f}(-x)\right| \le \frac{1}{2}p^2x^2 \le \frac{1}{2}px$, and so $(1-x)^p \ge 1 - px - \frac{1}{2}px = 1 - \frac{3}{2}px$.

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited

For $u \geq 0$, we need to be more detailed. We get $\max_{\xi \in \mathcal{I}}(1+\xi)^{p-2} = (1+x)^{p-2}$. Since $x \leq \frac{1}{p} \leq \frac{1}{p-2}$, we get $p < 2 + \frac{1}{x}$, so that $(1+x)^{p-2} \leq (1+x)^{\frac{1}{x}} \leq 3$. So,

$$|(1+x)^p - (1+px)| \leq \frac{1}{2}p^2x^2(1+x)^{p-2} \leq \frac{3}{2}p^2x^2 \leq \frac{3}{2}px$$

And therefore

$$(1+x)^p \leq 1 + px + \frac{3}{2}px \leq 1 + 3px$$

□

LEMMA E.5. *Suppose we sample n_0 times $s_1, \ldots, s_{n_0}$ uniformly from $[-1, 1]$. Then,* $\max_i v(s_i) \leq \frac{d+1}{\pi}\sqrt{\frac{n_0 \ln(2)}{\ln(\frac{1}{1-\delta})}} = \Theta(\frac{d\sqrt{n_0}}{\sqrt{\delta}})$ *with probability* $1 - \delta$.

Proof.

$$\begin{aligned}
\Pr[\max_i |v(s_i)| \leq F] &= (\Pr[v(s_1) \leq F])^{n_0} \\
\text{(Inverse Function of } v(s)) \qquad &= \left(\Pr\left[s_i \in \pm\sqrt{1 - \tfrac{(d+1)^2}{\pi^2 F^2}}\right]\right)^{n_0} \\
&= \left(1 - \left(\frac{d+1}{\pi F}\right)^2\right)^{n_0/2} \\
\text{(Let } x := (\tfrac{\pi F}{d+1})^2) \qquad &= \left(1 - \frac{1}{x}\right)^{x \cdot \frac{n_0}{2x}} \\
(x \geq 2) \qquad &\geq 0.25^{\frac{n_0}{2x}} \\
(0.25 = 2^{-2}) \qquad &= 2^{\frac{-n_0}{x}}
\end{aligned}$$

Making this fail with probability δ, we get

$$\begin{aligned}
2^{\frac{-n_0}{x}} &\geq 1 - \delta \\
-\frac{n_0}{x}\ln(2) &\geq \ln(1-\delta) \\
x &\leq -\frac{n_0 \ln(2)}{\ln(1-\delta)} \\
(\frac{\pi F}{d+1})^2 &\leq \frac{n_0 \ln(2)}{\ln(\frac{1}{1-\delta})} \\
F &\leq \frac{d+1}{\pi}\sqrt{\frac{n_0 \ln(2)}{\ln(\frac{1}{1-\delta})}}
\end{aligned}$$

□

Downloaded 06/17/23 to 73.222.54.66 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © 2023 by SIAM
Unauthorized reproduction of this article is prohibited