Chebyshev Sampling is Optimal for $L_p$ Polynomial Regression

Raphael A. Meyer

New York University

Tandon School of Engineering
Outline of Talk

1. Background
   - Problem Statement
   - Prior Work
   - Open Needs

2. Our Results
   - Upper Bounds
   - Lower Bounds

3. Our Techniques
   - From Lewis Weights to Jacobi Polynomials
   - Plenty not discussed here
We want to fit a function $f : [-1, 1] \to \mathbb{R}$ with a degree $d$ polynomial $\hat{q}$.
We want to fit a function $f : [-1, 1] \rightarrow \mathbb{R}$ with a degree $d$ polynomial $\hat{q}$. We can observe $f(t)$ at any $t \in [-1, 1]$. 
We want to fit a function $f : [-1, 1] \rightarrow \mathbb{R}$ with a degree $d$ polynomial $\hat{q}$.

We can observe $f(t)$ at any $t \in [-1, 1]$.

Goal: find polynomial $\hat{q}$ to minimize $L_p$ error:

$$\| f - \hat{q} \|_p^p \leq (1 + \varepsilon) \min_{\text{degree}(q)=d} \| f - \hat{q} \|_p^p$$

where $\| f \|_p := \int_{-1}^{1} |f(t)|^p dt$
We want to fit a function $f : [-1, 1] \rightarrow \mathbb{R}$ with a degree $d$ polynomial $\hat{q}$.

We can observe $f(t)$ at any $t \in [-1, 1]$.

Goal: find polynomial $\hat{q}$ to minimize $L_p$ error:

$$\|f - \hat{q}\|_p \leq (1 + \varepsilon) \min_{\text{degree}(q) = d} \|f - \hat{q}\|_p$$

where $\|f\|_p := \int_{-1}^{1} |f(t)|^p dt$
We want to fit a function $f : [-1, 1] \rightarrow \mathbb{R}$ with a degree $d$ polynomial $\hat{q}$. We can observe $f(t)$ at any $t \in [-1, 1]$.

Goal: find polynomial $\hat{q}$ to minimize $L_p$ error:

$$\|f - \hat{q}\|^p_p \leq (1 + \varepsilon) \min_{\text{degree}(q) = d} \|f - \hat{q}\|^p_p$$

where $\|f\|^p_p := \int_{-1}^{1} |f(t)|^p dt$
We want to fit a function $f : [-1, 1] \to \mathbb{R}$ with a degree $d$ polynomial $\hat{q}$.

We can observe $f(t)$ at any $t \in [-1, 1]$.

Goal: find polynomial $\hat{q}$ to minimize $L_p$ error:

$$\|f - \hat{q}\|_p^p \leq (1 + \varepsilon) \min_{\text{degree}(q)=d} \|f - \hat{q}\|_p^p$$

where $\|f\|_p^p := \int_{-1}^{1} |f(t)|^p dt$
Given: query access to $f$, maximum degree $d$, parameter $p$
Return: polynomial approximation $\hat{q}$

Two big questions:

1. How many observations are necessary?
   - If $f$ is a degree-$d$ polynomial, $n = \Omega(d)$ is needed
   - Larger $p$ needs more observations

2. How should we pick our observations?
   - Uniform sampling uses $n = O(d^2)$ queries
The Big Questions

Given: query access to $f$, maximum degree $d$, parameter $p$
Return: polynomial approximation $\hat{q}$

Two big questions:

1. How many observations are necessary? **Answer:** $n = \tilde{O}(dp^4)$ suffices
   - If $f$ is a degree-$d$ polynomial, $n = \Omega(d)$ is needed
   - Larger $p$ needs more observations

2. How should we pick our observations?
   - Uniform sampling uses $n = O(d^2)$ queries
Given: query access to $f$, maximum degree $d$, parameter $p$
Return: polynomial approximation $\hat{q}$

Two big questions:

1. How many observations are necessary? **Answer:** $n = \tilde{O}(dp^4)$ suffices
   - If $f$ is a degree-$d$ polynomial, $n = \Omega(d)$ is needed
   - Larger $p$ needs more observations

2. How should we pick our observations? **Answer:** Chebyshev Sampling
   - Uniform sampling uses $n = O(d^2)$ queries
Prior Work\(^1\) says:

For \(p = 2, \infty\), draw \(n = \tilde{O}(d)\) iid samples with PDF 
\[
v(t) := \frac{1}{\pi \sqrt{1 - t^2}}
\]

Then solve a Vandermonde matrix \(\ell_p\) regression problem.

\(^1\)Price Chen 2019, [Kane Karmalkar Price 2017]
Prior Work\textsuperscript{1} says:

For \( p = 2, \infty \), draw \( n = \tilde{O}(d) \) iid samples with PDF \( v(t) := \frac{1}{\pi \sqrt{1-t^2}} \)

Then solve a Vandermonde matrix \( \ell_p \) regression problem.

We show this works for all \( p \geq 1, d \geq 1, \varepsilon > 0 \)

\textsuperscript{1}[Price Chen 2019], [Kane Karmalkar Price 2017]
Our Contributions

Given: query access to $f$, maximum degree $d$, parameter $p$

**Algorithm** Chebyshev sampling for $L_p$ polynomial approximation

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 queries $b_i := f(t_i)$ for all $i \in [n]$
3. Build $A, S$ with $[A]_{i,j} = t_j^{i-1}$ and $[S]_{ii} = \left(\frac{d}{np} \sqrt{1 - t_i^2}\right)^{1/p}$
4. Compute $x = \arg\min_{x \in \mathbb{R}^{d+1}} \|SAx - Sb\|_p$
5. Return $q(t) = \sum_{i=0}^{d} x_i t^i$

Subtlety: for non-constant $\varepsilon$, $n = \tilde{O}\left(\frac{dp^4}{\varepsilon^{2p+2}}\right)$, run above algorithm twice
Chebyshev Sampling is Optimal for $L_p$ Polynomial Regression

Raphael A. Meyer

New York University
Tandon School of Engineering
Reinterpret the problem as $\ell_p$ regression with an “infinitely tall matrix”:

$$\min_{\deg(q) \leq d} \|q - f\|_p = \min_{x \in \mathbb{R}^{d+1}} \|P x - f\|_p$$

“Columns” of $P$ are monomials, “Rows” of $P$ are $[1 \ t \ t^2 \ \ldots \ t^d]$. 

Generalize prior work on Row-Sampling for $\ell_p$ Matrix Regression

---

$^2$[Chen et al. 2016], [Price Chen 2019], [Avron et al. 2019], [Meyer Musco 2020], ...
Leverage Function Prior Work for $p = 2$

For tall-and-skinny matrix $A \in \mathbb{R}^{n \times d}$, the Leverage Score for Row $i$ is

$$\tau[A](i) := \max_x \frac{[Ax]^2_i}{\|Ax\|^2_2}$$

With three key properties:
For tall-and-skinny matrix $A \in \mathbb{R}^{n \times d}$, the Leverage Score for Row $i$ is

$$\tau[A](i) := \max_x \frac{[Ax]^2_i}{\|Ax\|^2_2}$$

With three key properties:

1. Sampling $\tilde{O}(d)$ from $A$ rows preserves Least-Squares ($p = 2$) error
For tall-and-skinny matrix $A \in \mathbb{R}^{n \times d}$, the Leverage Score for Row $i$ is

$$\tau[A](i) := \max_x \frac{(Ax)^2_i}{\|Ax\|_2^2}$$

With three key properties:

1. Sampling $\tilde{O}(d)$ from $A$ rows preserves Least-Squares ($p = 2$) error
2. For any change-of-basis $B \in \mathbb{R}^{d \times d}$, we have $\tau[AB](i) = \tau[A](i)$
For tall-and-skinny matrix $A \in \mathbb{R}^{n \times d}$, the Leverage Score for Row $i$ is

$$\tau[A](i) := \max_x \frac{(Ax)^2_i}{\|Ax\|_2^2}$$

With three key properties:

1. Sampling $\tilde{O}(d)$ from $A$ rows preserves Least-Squares ($p = 2$) error
2. For any change-of-basis $B \in \mathbb{R}^{d \times d}$, we have $\tau[AB](i) = \tau[A](i)$
3. If $A$ has orthonormal columns, then $\tau[A](i) = \|a_i\|_2^2$ are row-norms
For tall-and-skinny matrix $A \in \mathbb{R}^{n \times d}$, the Leverage Score for Row $i$ is

$$\tau[A](i) := \max_x \frac{[Ax]_i^2}{\|Ax\|^2}$$

With three key properties:

1. Sampling $\tilde{O}(d)$ from $A$ rows preserves Least-Squares ($p = 2$) error
2. For any change-of-basis $B \in \mathbb{R}^{d \times d}$, we have $\tau[AB](i) = \tau[A](i)$
3. If $A$ has orthonormal columns, then $\tau[A](i) = \|a_i\|^2_2$ are row-norms

So, for operators instead of matrices,

Define Leverage Function at time $t$:

$$\tau[P](t) := \max_x \frac{(Px(t))^2}{\|Px\|^2}$$

Which has the same 3 properties
Question: How can we bound \( \tau[\mathcal{P}](t) \leq d \frac{1}{\pi \sqrt{1-t^2}} \)?
Question: How can we bound $\tau[\mathcal{P}](t) \leq d \frac{1}{\pi \sqrt{1-t^2}}$?

Change the basis of $\mathcal{P}$ to have Legendre Polynomials as columns:

$$
\int_{-1}^{1} L_i(t) L_j(t) \, dt = 1_{[i=j]}
$$
Question: How can we bound $\tau[\mathcal{P}](t) \leq d \frac{1}{\pi \sqrt{1-t^2}}$?

Change the basis of $\mathcal{P}$ to have Legendre Polynomials as columns:

$$\int_{-1}^{1} L_i(t) L_j(t) \, dt = \mathbf{1}_{[i=j]}$$

Then, by Uniform Bounds on Legendre Polynomials [Lorch 1983],

$$\tau[\mathcal{P}](t) = \sum_{i=0}^{d} (L_i(t))^2 \leq 2d \frac{1}{\pi \sqrt{1-t^2}}$$
Lewis Weights\(^4\) Now \(p \geq 1\)

For matrix \(A \in \mathbb{R}^{n \times d}\), weights \(w_1, \ldots, w_n\) are \(\ell_p\) Lewis Weights of \(A\) if

\[
\tau[W^{\frac{1}{2} - \frac{1}{p}} A](i) = w_i
\]

where \([W]_{ii} = w_i\) is a diagonal matrix.

---

\(^3\) [Cohen Peng 2015], [Musco et al. 2022]

\(^4\) [Cohen Peng 2015], [Musco et al. 2022]
For matrix $A \in \mathbb{R}^{n \times d}$, weights $w_1, \ldots, w_n$ are $\ell_p$-Lewis Weights of $A$ if

$$\tau[W^{\frac{1}{2}} - \frac{1}{p} A](i) = w_i$$

where $[W]_{ii} = w_i$ is a diagonal matrix.

1. Guess-and-check definition
Lewis Weights\textsuperscript{4} Now $p \geq 1$

For matrix $A \in \mathbb{R}^{n \times d}$, weights $w_1, \ldots, w_n$ are \textit{$\ell_p$ Lewis Weights} of $A$ if

$$\tau[W^{\frac{1}{2} - \frac{1}{p}} A](i) = w_i$$

where $[W]_{ii} = w_i$ is a diagonal matrix.

1. Guess-and-check definition
2. Sampling $\tilde{O}(d^{p/2})$ rows wrt $\ell_p$ Lewis weights preserves $\ell_p$ regression error

\textsuperscript{3}[Meyer et al 2022]
\textsuperscript{4}[Cohen Peng 2015], [Musco et al. 2022]

© New York University
Lewis Weights \(^4\)  Now \(p \geq 1\)

For matrix \(A \in \mathbb{R}^{n \times d}\), weights \(w_1, \ldots, w_n\) are \(\ell_p\) Lewis Weights of \(A\) if

\[
\tau [W^{\frac{1}{2}} - \frac{1}{p} A](i) = w_i
\]

where \([W]_{ii} = w_i\) is a diagonal matrix.

1. Guess-and-check definition
2. Sampling \(\tilde{O}(dp^2)\) rows wrt \(\ell_p\) Lewis weights preserves \(\ell_p\) regression error

---

\(^3\) [Meyer et al 2022]
\(^4\) [Cohen Peng 2015], [Musco et al. 2022]
For matrix $A \in \mathbb{R}^{n \times d}$, weights $w_1, \ldots, w_n$ are $\ell_p$-Lewis Weights of $A$ if

$$\tau[W^{\frac{1}{2} - \frac{1}{p}} A](i) = w_i$$

where $[W]_{ii} = w_i$ is a diagonal matrix.

1. Guess-and-check definition
2. Sampling $\tilde{O}(dp^2)$ rows wrt $\ell_p$ Lewis weights preserves $\ell_p$ regression error

Weaker goalpost: it’s enough to sample by $w_1, \ldots, w_n$ with

$$\frac{1}{C} w_i \leq \tau[W^{\frac{1}{2} - \frac{1}{p}} A](i) \leq C w_i \quad \text{for all } i \in [n]$$

---

3 [Meyer et al 2022]
4 [Cohen Peng 2015], [Musco et al. 2022]
For matrix $A \in \mathbb{R}^{n \times d}$, weights $w_1, \ldots, w_n$ are $\ell_p$ Lewis Weights of $A$ if

$$\tau[W^{\frac{1}{2} - \frac{1}{p}} A](i) = w_i$$

where $[W]_{ii} = w_i$ is a diagonal matrix.

1. Guess-and-check definition
2. Sampling $\tilde{O}(dp^2)$ rows wrt $\ell_p$ Lewis weights preserves $\ell_p$ regression error

Weaker goalpost: it’s enough to sample by $w_1, \ldots, w_n$ with

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

---

3 [Meyer et al 2022]
4 [Cohen Peng 2015], [Musco et al. 2022]
Behold Orthogonal Polynomials

Now \( p \geq 1 \)

Idea: Guess \( v(t) = d \frac{1}{\pi \sqrt{1-t^2}} \) are Lewis Weights
Now $p \geq 1$

Idea: Guess $v(t) = d \frac{1}{\pi \sqrt{1-t^2}}$ are Lewis Weights

Change the basis of $\mathcal{P}$ to have **Gegenbauer Polynomials** as columns:

$$
\int_{-1}^{1} \frac{J_i^{(\alpha)}(t) J_j^{(\alpha)}(t) (1-t^2)^{\alpha - \frac{1}{2}}}{\pi \sqrt{1-t^2}} dt = \mathbb{I}_{[i=j]} 
$$
Behold Orthogonal Polynomials Now $p \geq 1$

Idea: Guess $v(t) = d \frac{1}{\pi \sqrt{1-t^2}}$ are Lewis Weights

Change the basis of $\mathcal{P}$ to have Gegenbauer Polynomials as columns:

$$\int_{-1}^{1} J_i^{(\alpha)}(t) J_j^{(\alpha)}(t) (1-t^2)^{\alpha-\frac{1}{2}} \, dt = \mathbb{1}_{[i=j]}$$

Then $\mathcal{V}^{\frac{1}{2}} \frac{1}{p} \mathcal{P}$ has orthonormal columns, so by [Nevai et al. 1997]

$$\tau[\mathcal{V}^{\frac{1}{2}} \frac{1}{p} \mathcal{P}](t) = (1-t^2)^{\frac{1}{p}-\frac{1}{2}} \sum_{i=0}^{d} (J_i^{(\alpha)}(t))^2 \leq Cd \frac{1}{\pi \sqrt{1-t^2}}$$
We’re not done yet

We need to prove \( \frac{1}{C} v(t) \leq \tau [\mathcal{V}^{1/2} - \frac{1}{p} \mathcal{P}] (t) \leq C v(t) \) for all \( t \in [-1, 1] \).
We're not done yet

We need to prove \( \frac{1}{C} v(t) \leq \tau \left[ \mathcal{V}^{\frac{1}{2}} - \frac{1}{p} \mathcal{P} \right](t) \leq C v(t) \) for all \( t \in [-1, 1] \).

For \( p = 1 \),

\[
\frac{\tau [\mathcal{V}^{-\frac{1}{2}} \mathcal{P}](t)}{v(t)} = 1 + \frac{1 - U_{2(d+1)}(t)}{2(d + 1)} \to 0 \quad \text{as} \ t \to \pm 1
\]
We’re not done yet

Refined Analysis for $t \to 1$ via “Clipped Chebyshev Measure”

Matrix Guarantees Extend to Operators via “Two-Stage Sampling”
Given: query access to $f$, maximum degree $d$, parameter $p$

Return: polynomial approximation $\hat{q}$

Two big questions:

1. How many observations are necessary?
   - If $f$ is a degree-$d$ polynomial, $n = \Omega(d)$ is needed
   - Larger $p$ needs more observations

2. How should we pick our observations?
   - Uniform sampling uses $n = O(d^2)$ queries

Main Analysis that I Presented:

- Define Operator Lewis Weights
- Relate Operator Lewis Weights to Gegenbauer Polynomials
- Prior work relates Gegenbauer Polynomials to Chebyshev measure
- So much not explained here....
Given: query access to $f$, maximum degree $d$, parameter $p$
Return: polynomial approximation $\hat{q}$

Two big questions:

1. How many observations are necessary? Answer: $n = \tilde{O}(dp^4)$ suffices
   - If $f$ is a degree-$d$ polynomial, $n = \Omega(d)$ is needed
   - Larger $p$ needs more observations

2. How should we pick our observations?
   - Uniform sampling uses $n = O(d^2)$ queries

Main Analysis that I Presented:

- Define Operator Lewis Weights
- Relate Operator Lewis Weights to Gegenbauer Polynomials
- Prior work relates Gegenbauer Polynomials to Chebyshev measure
- So much not explained here....
Summary

Given: query access to $f$, maximum degree $d$, parameter $p$
Return: polynomial approximation $\hat{q}$

Two big questions:

1. How many observations are necessary? **Answer:** $n = \tilde{O}(dp^4)$ suffices
   - If $f$ is a degree-$d$ polynomial, $n = \Omega(d)$ is needed
   - Larger $p$ needs more observations

2. How should we pick our observations? **Answer:** Chebyshev Sampling
   - Uniform sampling uses $n = O(d^2)$ queries

Main Analysis that I Presented:

- Define Operator Lewis Weights
- Relate Operator Lewis Weights to Gegenbauer Polynomials
- Prior work relates Gegenbauer Polynomials to Chebyshev measure
- So much not explained here....