Leverage scores, Christoffel functions, and Applications of RandNLA beyond NLA.

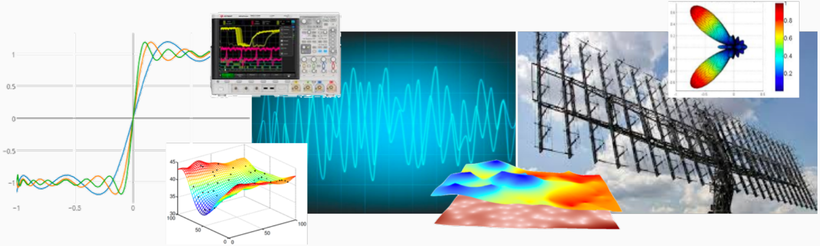
Christopher Musco, New York University

TODAY'S TOPIC

Goal: Expand the influence of RandNLA to other pillars of applied and computational mathematics.

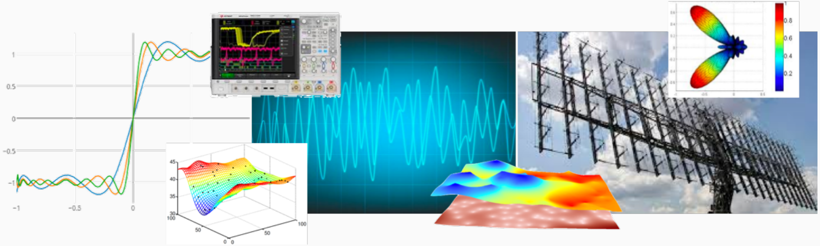
Potential for ideas from randomized linear algebra to impact:

Approximation theory, signal processing, wireless communications, spatial statistics, etc.



Potential for ideas from randomized linear algebra to impact:

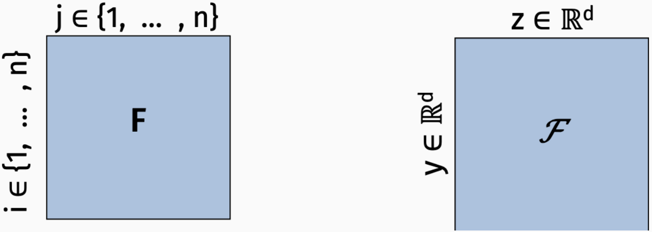
Approximation theory, signal processing, wireless communications, spatial statistics, etc.



AKA: What I learned between last year at Simons and now.

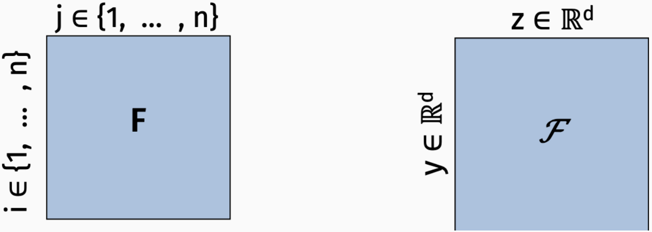
Powerful tool: Extend RandNLA sketching and sampling methods from matrices to continuous linear operators.

$$[\mathbf{F}\mathbf{x}]_i = \sum_j \mathbf{F}_{i,j}\mathbf{x}_j \quad \Longrightarrow \quad [\mathcal{F}x](y) = \int_{z \in S} \mathcal{F}(y,z)x(z)dz$$



Powerful tool: Extend RandNLA sketching and sampling methods from matrices to continuous linear operators.

$$[\mathbf{F}\mathbf{x}]_i = \sum_j \mathbf{F}_{i,j}\mathbf{x}_j \quad \Longrightarrow \quad [\mathcal{F}x](y) = \int_{z \in S} \mathcal{F}(y,z)x(z)dz$$

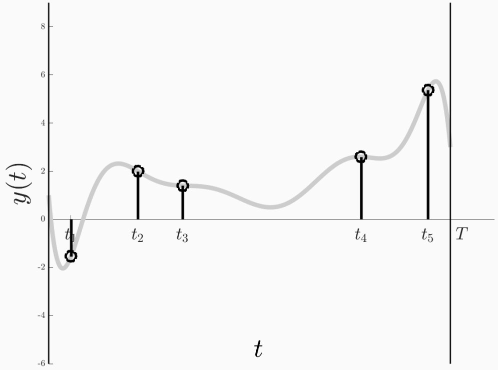


Randomized Functional Analysis.

- Continuous function interpolation (polynomials, bandlimited, sparse Fourier, etc.)
- Kernel approximation in machine learning (random Fourier features methods).
- Toeplitz covariance approximation + other signal processing problem.
- What else?

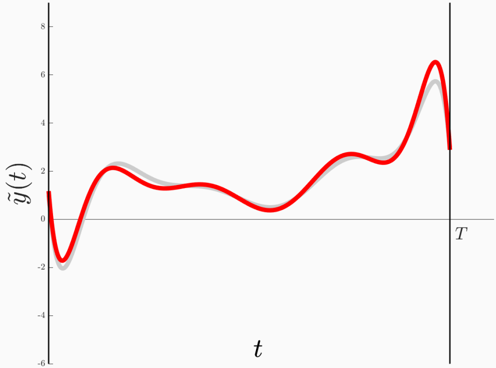
Polynomial regression:

- **Given:** Pointwise access to $f(x) : [0, T] \to \mathbb{R}$ for For $x \in [0, T]$.



Polynomial regression:

- **Given:** Pointwise access to $f(x) : [0, T] \to \mathbb{R}$ for For $x \in [0, T]$.
- **Goal:** Find good degree q polynomial fit to f on $[0, T]$.



Approximation goal:

If there exists a degree q polynomial p^* with $\|p^* - f\| \leq \epsilon$, return a degree q polynomial p with $\|p - f\| \leq O(\epsilon)$.

Approximation goal:

If there exists a degree q polynomial p^* with $\|p^* - f\| \leq \epsilon$, return a degree q polynomial p with $\|p - f\| \leq O(\epsilon)$.

- ∞-norm: $\|p - f\| = \max_{x \in [0,T]} |f(x) - p(x)|$.

Approximation goal:

If there exists a degree q polynomial p^* with $\|p^* - f\| \leq \epsilon$, return a degree q polynomial p with $\|p - f\| \leq O(\epsilon)$.

- ∞-norm: $\|p - f\| = \max_{x \in [0,T]} |f(x) - p(x)|$.
- $L2$-norm: $\|p - f\| = \int_0^T |f(x) - p(x)|^2 dx$.

Approximation goal:

If there exists a degree q polynomial p^* with $\|p^* - f\| \leq \epsilon$, return a degree q polynomial p with $\|p - f\| \leq O(\epsilon)$.

- ∞-norm: $\|p - f\| = \max_{x \in [0,T]} |f(x) - p(x)|$.
- $L2$-norm: $\|p - f\| = \int_0^T |f(x) - p(x)|^2 dx$.
- $L1$-norm: $\|p - f\| = \int_0^T |f(x) - p(x)| dx$.

Approximation goal:

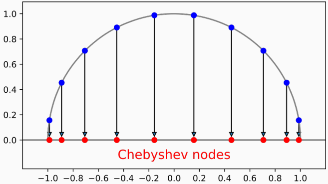
If there exists a degree q polynomial p^* with $\|p^* - f\| \leq \epsilon$, return a degree q polynomial p with $\|p - f\| \leq O(\epsilon)$.

- ∞-norm: $\|p - f\| = \max_{x\in[0,T]} |f(x) - p(x)|$.
- $L2$-norm: $\|p - f\| = \int_0^T |f(x) - p(x)|^2 dx$.
- $L1$-norm: $\|p - f\| = \int_0^T |f(x) - p(x)| dx$.
- ...

∞-norm: $\|p - f\|_\infty = \max_{x\in[0,T]} |f(x) - p(x)|$.

$$\infty\text{-norm: } \|p - f\|_\infty = \max_{x\in[0,T]} |f(x) - p(x)|.$$

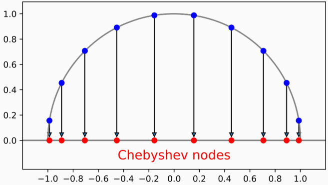
1. Take samples at $q + 1$ Chebyshev nodes.



2. Interpolation or Chebyshev expansion to find p.

$$\infty\text{-norm: } \|p - f\|_\infty = \max_{x\in[0,T]} |f(x) - p(x)|.$$

1. Take samples at $q + 1$ Chebyshev nodes.



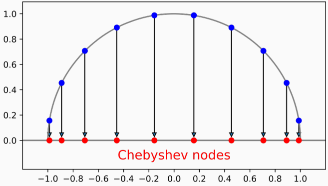
2. Interpolation or Chebyshev expansion to find p.

Guarantee: $\|p - f\|_\infty \leq O(\log k) \cdot \|p^* - f\|_\infty$.

"On the maximum errors of polynomial approximations defined by interpolation and by least squares criteria." [Powell, 1967].

∞-norm: $\|p - f\|_\infty = \max_{x\in[0,T]} |f(x) - p(x)|$.

1. Take samples at $q + 1$ Chebyshev nodes.

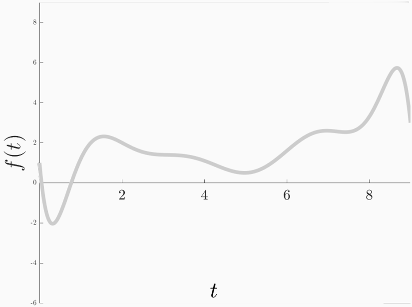


2. Interpolation or Chebyshev expansion to find p.

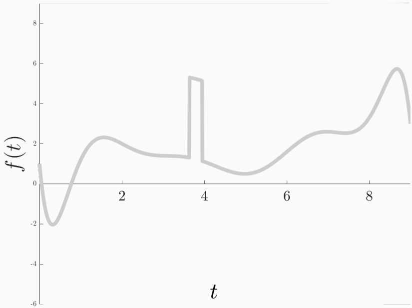
Guarantee: $\|p - f\|_\infty \leq O(\log k) \cdot \|p^* - f\|_\infty$.

"On the maximum errors of polynomial approximations defined by interpolation and by least squares criteria." [Powell, 1967].

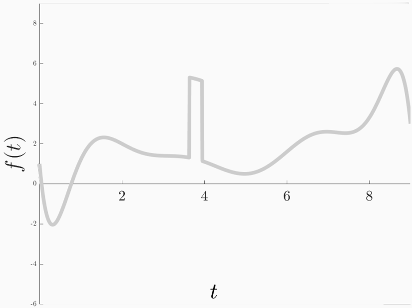
∞-norm bounds lack any robustness properties.



∞-norm bounds lack any robustness properties.

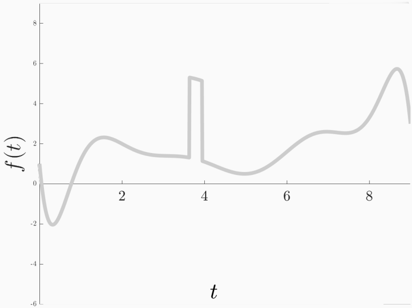


∞-norm bounds lack any robustness properties.



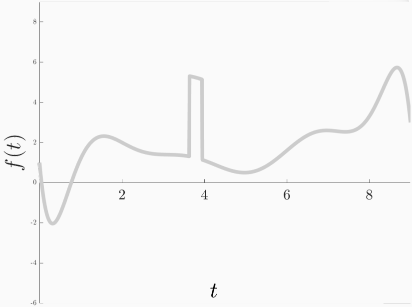
No p^* with small $\|f - p^*\|_\infty$, but we could still recover a good $L2$-norm or $L1$-norm approximation.

∞-norm bounds lack any robustness properties.

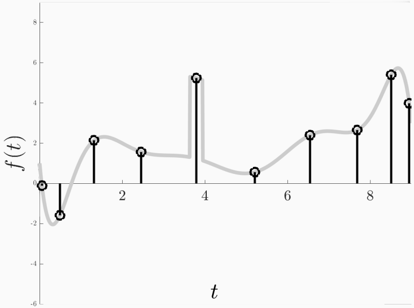


No p^* with small $\|f - p^*\|_\infty$, **but we could still recover a good $L2$-norm or $L1$-norm approximation.**

This is an issue with the methods, not just the analysis!



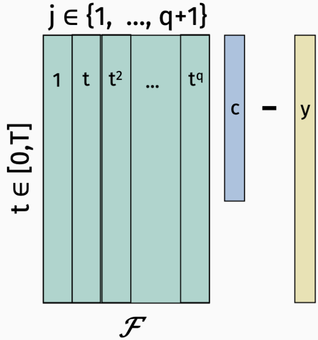
This is an issue with the methods, not just the analysis!



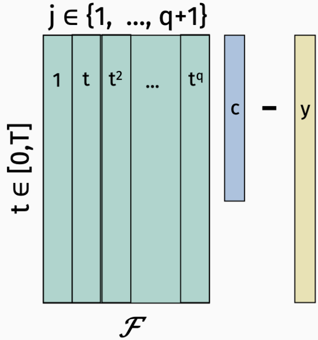
Randomness is <u>necessary</u> for robustness.

RandNLA approach: View as least-squares regression:

RandNLA approach: View as least-squares regression:



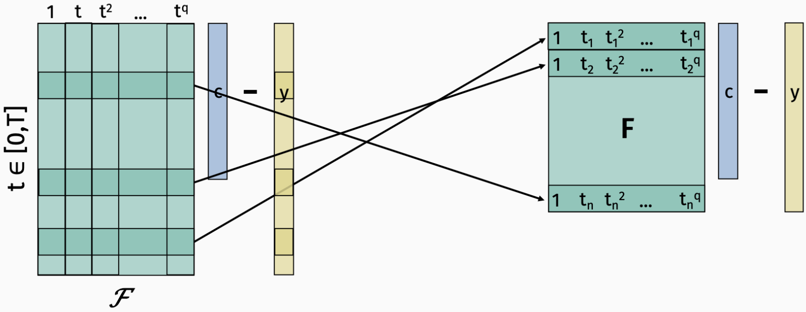
RandNLA approach: View as least-squares regression:



$$p_{\mathbf{c}}(t) = c_0 + c_1 t + \ldots + c_q t^q$$

$$\min_{\mathbf{c}} \|\mathcal{F}\mathbf{c} - \mathbf{y}\|_2 = \int_0^T |y(t) - p_{\mathbf{c}}(t)|^2 dt$$

Discretize via subsampling.



LEVERAGE SCORE SAMPLING

[Drineas, Mahoney, Muthukrishnan 2006], [Sarlos 2006], [Spielman, Srivastava 2008], etc.

Rank q operator $\implies$ suffices to sample $O(q \log q)$ rows from $\mathcal{F}$ with probability proportional to statistical leverage scores.

[Drineas, Mahoney, Muthukrishnan 2006], [Sarlos 2006], [Spielman, Srivastava 2008], etc.

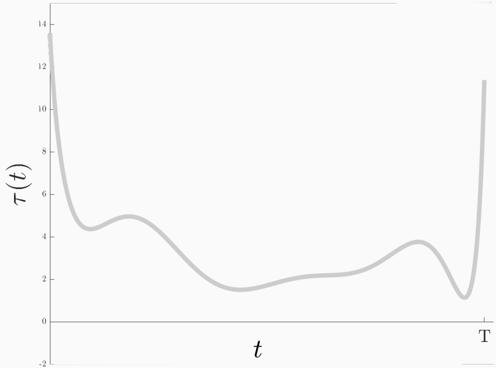
Rank q operator $\implies$ suffices to sample $\mathbf{O(q \log q)}$ rows from $\mathcal{F}$ with probability proportional to **statistical leverage scores**.

Definition (Leverage Score)

The leverage scores of row $\mathbf{a}_i$ in a matrix $\mathbf{A}$, $\tau(\mathbf{a}_i)$ equals:

$$\begin{aligned}\tau(\mathbf{a}_i) &= \mathbf{a}_i^T(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{a}_i \\ &= \min_{\mathbf{y}:\mathbf{A}^T\mathbf{y}=\mathbf{a}_i} \|\mathbf{y}\|_2^2 \\ &= \max_{\mathbf{x}} \frac{(\mathbf{A}\mathbf{x})_i^2}{\|\mathbf{A}\mathbf{x}\|_2^2}\end{aligned}$$

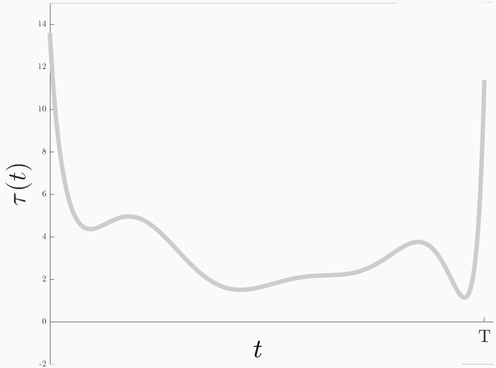
For the linear operator with continuous column space, we have a leverage function $\tau : [0, T] \to \mathbb{R}^+$.



For the linear operator with continuous column space, we have a leverage function $\tau : [0, T] \to \mathbb{R}^+$.

Forget about computing the leverage function...

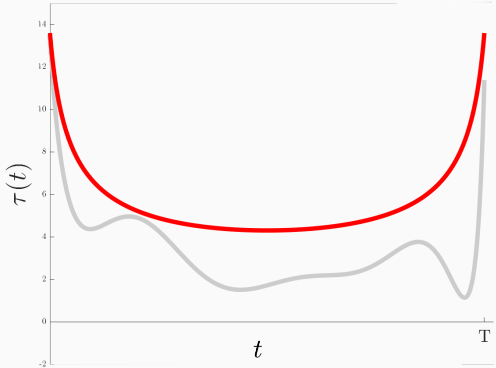
We need a closed form upper bound.



For the linear operator with continuous column space, we have a leverage function $\tau : [0, T] \rightarrow \mathbb{R}^{+}$.

Forget about computing the leverage function...

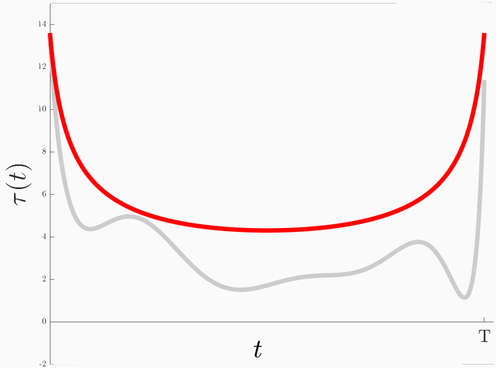
We need a closed form upper bound.



For the linear operator with continuous column space, we have a leverage function $\tau : [0, T] \to \mathbb{R}^+$.

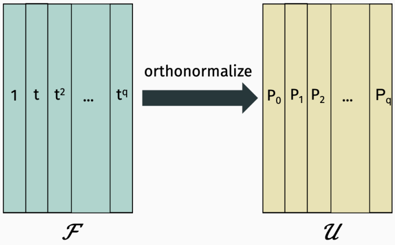
Forget about computing the leverage function...

We need a closed form upper bound.

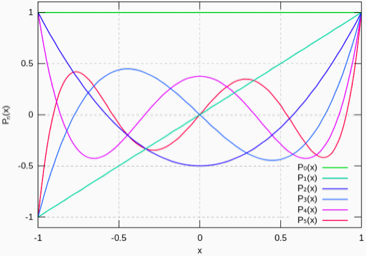


1) Row norms of any **orthonormal basis** for columns of $\mathcal{F}$.

1) Row norms of any **orthonormal basis** for columns of $\mathcal{F}$.



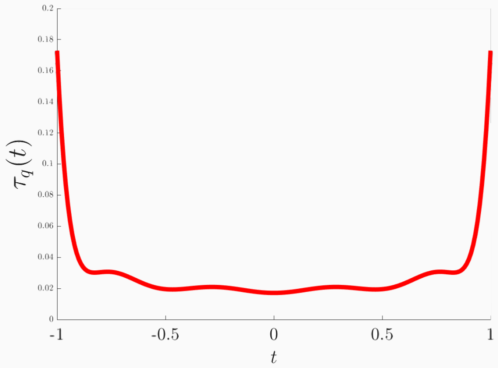
1) Row norms of any **orthonormal basis** for columns of $\mathcal{F}$.



Legendre Polynomials $P_0, P_1, \ldots, P_q$.

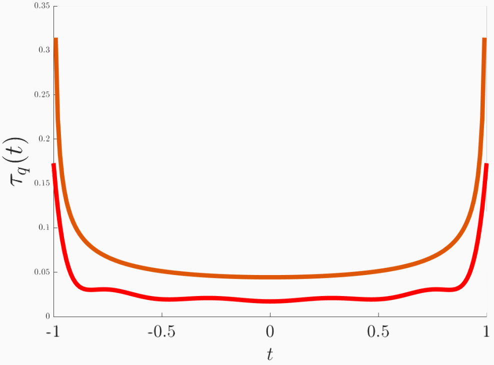
For all $i \neq j$, $\langle P_i, Pj \rangle = \int_{-1}^{1} P_i(t)P_j(t)dt = 0$.

1) Row norms of any **Legendre polynomial basis**.



$$\sum_{i=0}^{q} P_i(t)^2$$

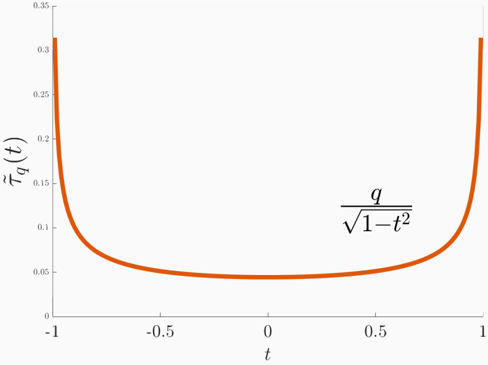
1) Row norms of any **Legendre polynomial basis**.



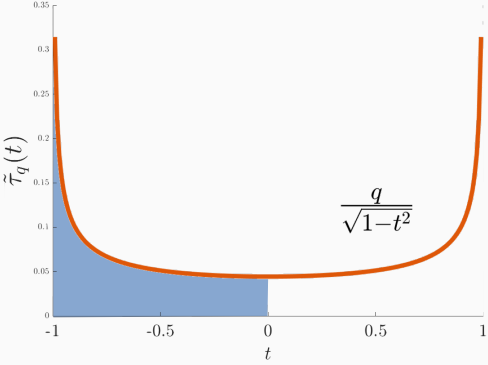
$$\sum_{i=0}^{q} P_i(t)^2 \leq \frac{cq}{\sqrt{1-t^2}}$$

[Lorch, '83]

Lorch's bound is tight up to constants:

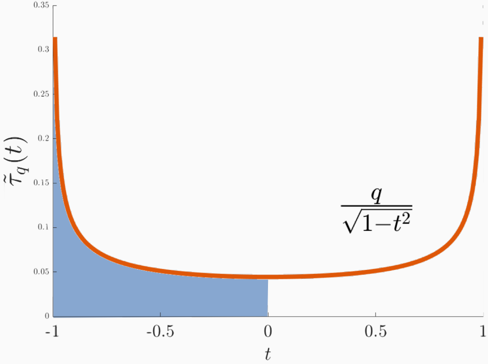


Lorch's bound is tight up to constants:



$$\int_{-1}^{1} \tilde{\tau}_q(t)dt = O(q).$$

Lorch's bound is tight up to constants:



$$\int_{-1}^{1} \tilde{\tau}_q(t)dt = O(q).$$

Only need to take $O(q \log q)$ samples.

Theorem (Cohen, Migliorati '16)

Suppose there exists a degree q polynomial p^* with $\|p^* - f\|_2 \leq \epsilon$ on $[0, T]$. Let $p = \arg\min \sum_{i=1}^{s} (f(t_i) - p(t_i))^2$ where $t_1, \ldots, t_s$ are sampled via the leverage score distribution for degree q polynomials. Then as long as $s = O(q \log q)$, $\|p - f\|_2 \leq O(\epsilon)$ with high probability.

Theorem (Cohen, Migliorati '16)

Suppose there exists a degree q polynomial p^* with $\|p^* - f\|_2 \leq \epsilon$ on $[0, T]$. Let $p = \arg\min \sum_{i=1}^{s}(f(t_i) - p(t_i))^2$ where $t_1, \ldots, t_s$ are sampled via the leverage score distribution for degree q polynomials. Then as long as $s = O(q \log q)$, $\|p - f\|_2 \leq O(\epsilon)$ with high probability.

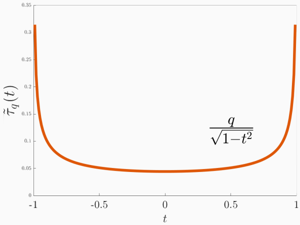
First near sample optimal robust polynomial interpolation via tools from RandNLA/matrix concentration!

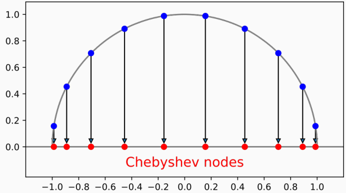
Theorem (Cohen, Migliorati '16)

Suppose there exists a degree q polynomial p^* with $\|p^* - f\|_2 \leq \epsilon$ on $[0, T]$. Let $p = \arg\min \sum_{i=1}^{s}(f(t_i) - p(t_i))^2$ where $t_1, \ldots, t_s$ are sampled via the leverage score distribution for degree q polynomials. Then as long as $s = O(q \log q)$, $\|p - f\|_2 \leq O(\epsilon)$ with high probability.

First near sample optimal robust polynomial interpolation via tools from RandNLA/matrix concentration!

See [Price, Chen COLT '19] for improvement to $O(q)$ samples, using ideas from Batson, Spielman, Srivastava '08].





Samples taken according to the same asymptotic density as the Chebyshev nodes.

Definition (Christoffel Function)

$$\lambda_q(t) = \sup_{\text{degree } q \text{ poly. } p} \frac{\int_{-1}^{1} p(t)^2 dt}{p(t)^2}.$$

Cohen and Migliorati sample proportional to $\frac{1}{\lambda_q(t)}$.

Definition (Christoffel Function)

$$\lambda_q(t) = \sup_{\text{degree } q \text{ poly. } p} \frac{\int_{-1}^{1} p(t)^2 dt}{p(t)^2}.$$

Cohen and Migliorati sample proportional to $\frac{1}{\lambda_q(t)}$.

Lots of other application in approximation theory: Jackson type theorems, understanding zeros of orthogonal polynomials, behavior or orthogonal expansion, etc.

Claim (Inverse Christoffel Function = Leverage Score)

$$\frac{1}{\lambda_q(t)} = \tau_q(t).$$

Claim (Inverse Christoffel Function = Leverage Score)

$$\frac{1}{\lambda_q(t)} = \tau_q(t).$$

Maximization characterization of leverage scores:

$$\tau(\mathbf{a}_i) = \max_{\mathbf{x}} \frac{(\mathbf{A}\mathbf{x})_i^2}{\|\mathbf{A}\mathbf{x}\|_2^2}$$

Claim (Inverse Christoffel Function = Leverage Score)

$$\frac{1}{\lambda_q(t)} = \tau_q(t).$$

Maximization characterization of leverage scores:

$$\tau(t) = \max_{\mathbf{x}} \frac{(\mathcal{F}\mathbf{x}(t))^2}{\|\mathcal{F}\mathbf{x}\|_2^2}$$

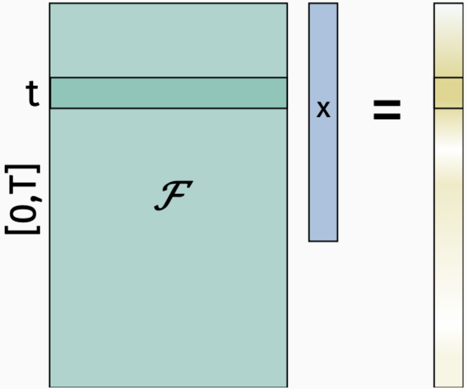
Claim (Inverse Christoffel Function = Leverage Score)

$$\frac{1}{\lambda_q(t)} = \tau_q(t).$$

Maximization characterization of leverage scores:

$$\tau(t) = \max_{\mathbf{x}} \frac{(\mathcal{F}\mathbf{x}(t))^2}{\|\mathcal{F}\mathbf{x}\|_2^2}$$

Follows from standard equation $\tau(\mathbf{a}_i) = \mathbf{a}_i^T(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{a}_i$ + Cauchy-Scharwz + basic linear algebra.



Is there a function in the range of $\mathcal{F}$ that is concentrated at t?

- Markov Brother's Inequality:

$$\tau_q(t) = \inf_{\text{degree } q \text{ poly. } p} \frac{p(t)^2}{\int_{-1}^{1} p(t)^2 dt} \leq O\left(q^2\right)$$

- Markov Brother's Inequality:

$$\tau_q(t) = \inf_{\text{degree } q \text{ poly. } p} \frac{p(t)^2}{\int_{-1}^{1} p(t)^2 dt} \leq O\left(q^2\right)$$

- Bernstein's Inequality:

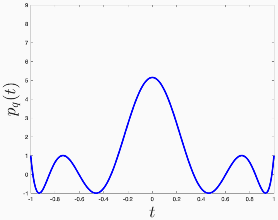
$$\tau_q(t) = \inf_{\text{degree } q \text{ poly. } p} \frac{p(t)^2}{\int_{-1}^{1} p(t)^2 dt} \leq O\left(\frac{q}{\sqrt{1-t^2}}\right)$$

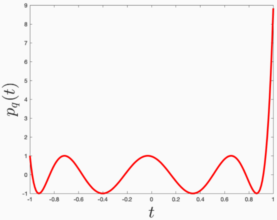
- **Markov Brother's Inequality**:

$$\tau_q(t) = \inf_{\text{degree } q \text{ poly. } p} \frac{p(t)^2}{\int_{-1}^{1} p(t)^2 dt} \leq O\left(q^2\right)$$

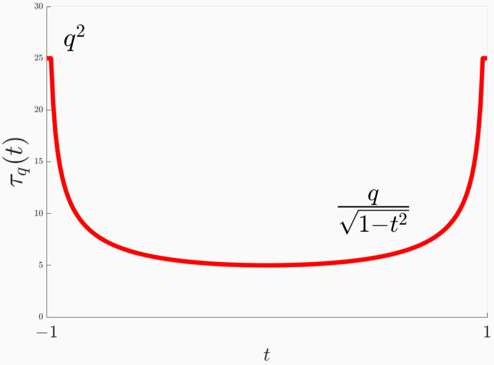
- **Bernstein's Inequality**:

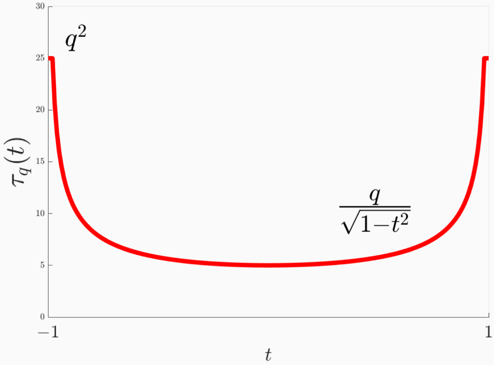
$$\tau_q(t) = \inf_{\text{degree } q \text{ poly. } p} \frac{p(t)^2}{\int_{-1}^{1} p(t)^2 dt} \leq O\left(\frac{q}{\sqrt{1-t^2}}\right)$$





Polynomials can "spike" more near the edge of an interval.

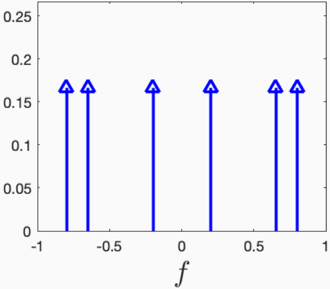

30
25
20
15
10
5
0
q^2
$\tau_q(t)$
$\frac{q}{\sqrt{1-t^2}}$
-1
1
t



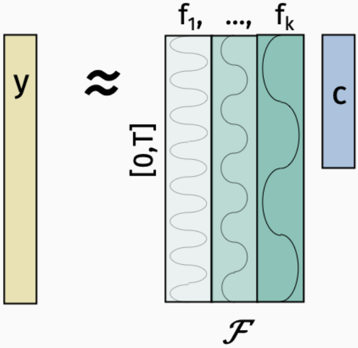
Additional implication: $O(q^2 \log q)$ uniformly sampled points are sufficient for robust polynomial fitting **[Cohen, Davenport, Leviatan, FoCM '13]**.

Anything beyond the polynomial operator?

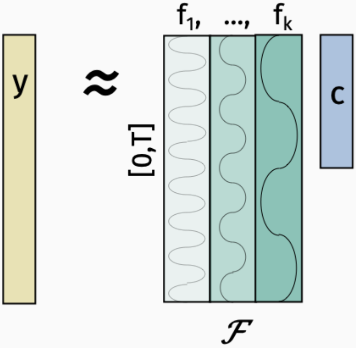
E.g. y is **Fourier sparse**. $\hat{y}(\xi)$ is supported on k frequencies.



Compressed sensing, applications in medical imaging, microscopy, RADAR, etc.



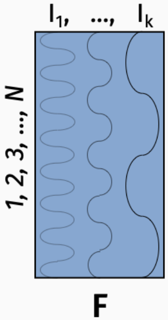
$$\mathcal{F}_j(t) = e^{-2\pi i f_j t}.$$



$$\mathcal{F}_j(t) = e^{-2\pi i f_j t}.$$

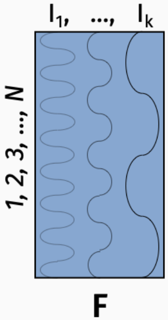
Nearly the same bounds holds for k-sparse Fourier operators, not matter what $f_1, \ldots, f_j$ are!

Such bounds are immediate for discrete Fourier matrices:



$$F_{j,k} = e^{-2\pi i l_j k}.$$

Such bounds are immediate for discrete Fourier matrices:



$$F_{j,k} = e^{-2\pi i l_j k}.$$

F is always a (scaled) orthonormal basis, so leverage scores are proportional to row norms, which are all the same.

[Halász 1983], [Kós, 2008], **[Erdélyi 2016]**, [Chen, Kane, Price, Song, FOCS 2016], [Chen, Price 2018], [Avron, Kapralov, Musco, Musco, Velingker, Zandieh, 2019]

[Halász 1983], [Kós, 2008], **[Erdélyi 2016]**, [Chen, Kane, Price, Song, FOCS 2016], [Chen, Price 2018], [Avron, Kapralov, Musco, Musco, Velingker, Zandieh, 2019]

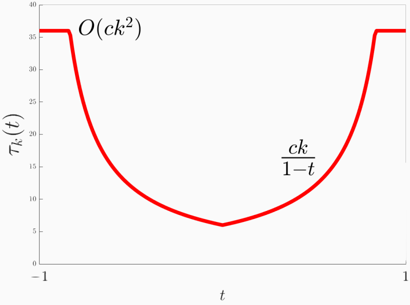
Theorem (Fourier sparse leverage)

When $\mathcal{F}$ is a k-sparse Fourier operator on $[0, T]$,

$$\frac{|\mathcal{F}x(t)|^2}{\|\mathcal{F}x\|_2^2} \leq ck^2$$

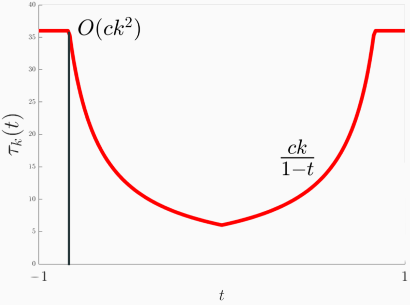
$$\frac{|\mathcal{F}x(t)|^2}{\|\mathcal{F}x\|_2^2} \leq ck/\min(t, T-t).$$

This upper bound is nearly tight:



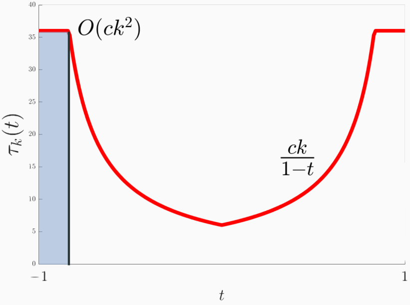
Total number of samples:

This upper bound is nearly tight:



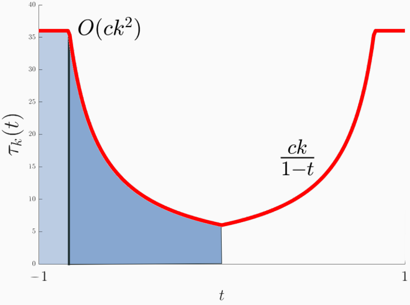
Total number of samples:

This upper bound is nearly tight:



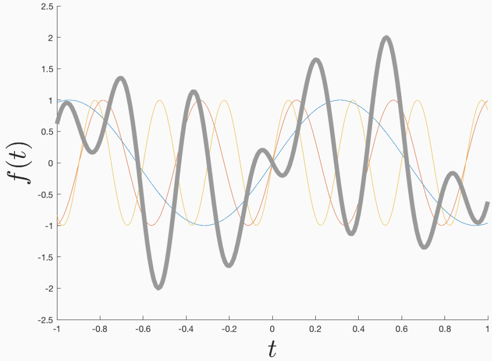
Total number of samples: k

This upper bound is nearly tight:



Total number of samples: $k + O(k \log k)$

Intuition: Sums of close frequencies look like modulated polynomials. Far frequencies are nearly orthogonal.



Theorem (Chen, Kane, Price, Song FOCS 2016)

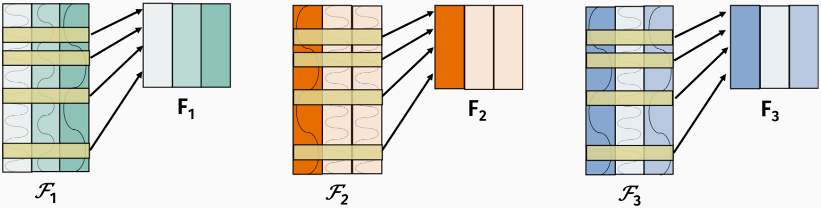
Given $y : [0, T] \to \mathbb{R}$. Suppose there is some k-sparse Fourier function $g(t) = \sum_{j=1}^{k} c_j e^{-2\pi i f_j t}$ with $\|y - g\|_2^2 \leq O(\epsilon)$. With $\tilde{O}(k^4)$ *samples, we can find a k-sparse Fourier function $\tilde{g}$ with:*

$$\|y - \tilde{g}\|_2^2 \leq O(\epsilon).$$

Theorem (Chen, Kane, Price, Song FOCS 2016)

Given $y : [0, T] \to \mathbb{R}$. Suppose there is some k-sparse Fourier function $g(t) = \sum_{j=1}^{k} c_j e^{-2\pi i f_j t}$ with $\|y - g\|_2^2 \leq O(\epsilon)$. With $\tilde{O}(k^4)$ samples, we can find a k-sparse Fourier function $\tilde{g}$ with:

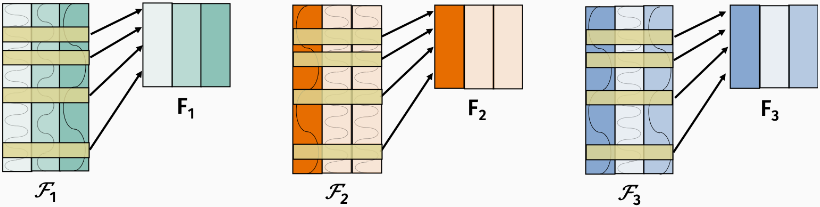
$$\|y - \tilde{g}\|_2^2 \leq O(\epsilon).$$



Theorem (Chen, Kane, Price, Song FOCS 2016)

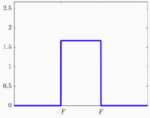
Given $y : [0, T] \to \mathbb{R}$. Suppose there is some k-sparse Fourier function $g(t) = \sum_{j=1}^{k} c_j e^{-2\pi i f_j t}$ with $\|y - g\|_2^2 \leq O(\epsilon)$. With $\tilde{O}(k^4)$ samples, we can find a k-sparse Fourier function $\tilde{g}$ with:

$$\|y - \tilde{g}\|_2^2 \leq O(\epsilon).$$

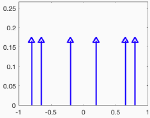


Major challenge: Can this be improved? To $O(k \log^c k)$?

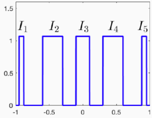
Implications for many other interpolation problems with Fourier structure:



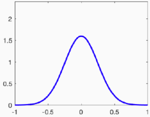
Bandlimited.



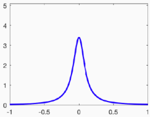
Sparse.



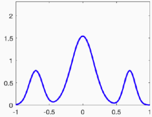
Multiband.



Gaussian.

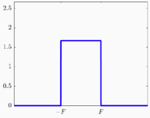


Cauchy-Lorentz.

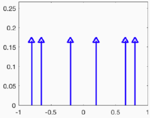


Gaussian mixture.

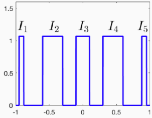
Implications for many other interpolation problems with Fourier structure:



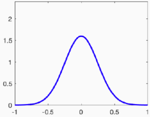
Bandlimited.



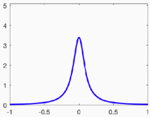
Sparse.



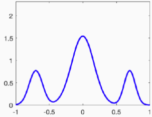
Multiband.



Gaussian.



Cauchy-Lorentz.

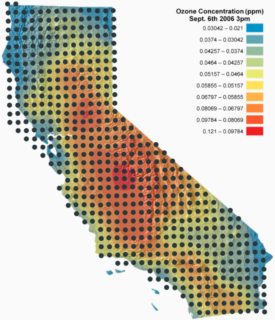


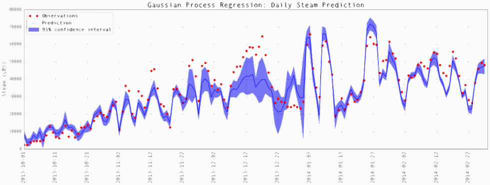
Gaussian mixture.

Applications in spectrum sensing/cognitive radio, medical imaging, time series, etc.

OTHER PRIORS

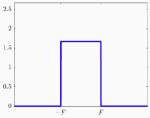
Smooth penalties underly Gaussian process regression, kriging, kernel ridge regression, etc.



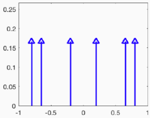


Countless applications in environmental science, geostatistics, image processing, economics, time series analysis, etc.

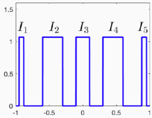
"A Universal Sampling Method for Reconstructing Signals with Simple Fourier Transforms" [AKMMVZ, STOC 2019].



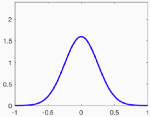
Bandlimited.



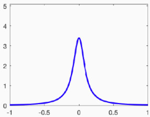
Sparse.



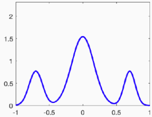
Multiband.



Gaussian.

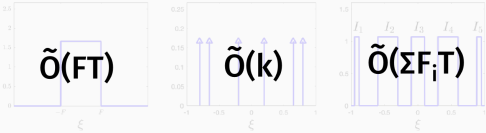


Cauchy-Lorentz.

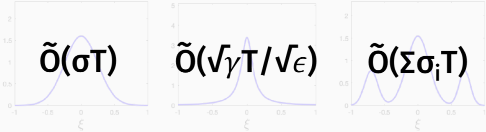


Gaussian mixture.

"A Universal Sampling Method for Reconstructing Signals with Simple Fourier Transforms" [AKMMVZ, STOC 2019].

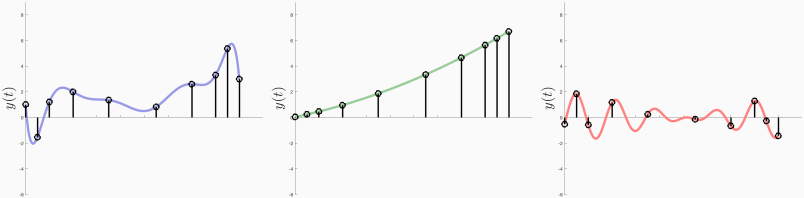


Bandlimited. Sparse. Multiband.



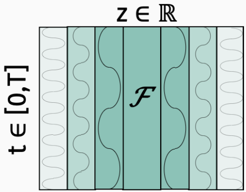
Gaussian. Cauchy-Lorentz. Gaussian mixture.

”A Universal Sampling Method for Reconstructing Signals with Simple Fourier Transforms” [AKMMVZ, STOC 2019].



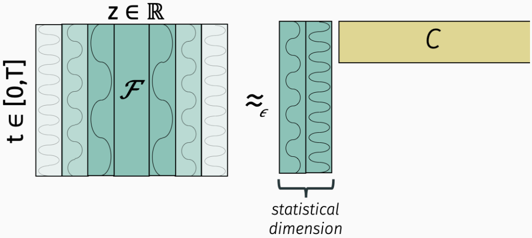
For these problems, we need to sample by an appropriately defined ridge leverage score:

$$\tau(t) = \max_g \frac{\frac{1}{T}|\mathcal{F}g(t)|^2}{\|\mathcal{F}g\|_2^2 + \epsilon\|g\|_2^2}$$

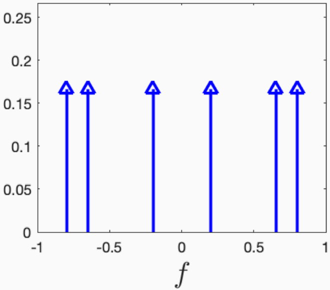


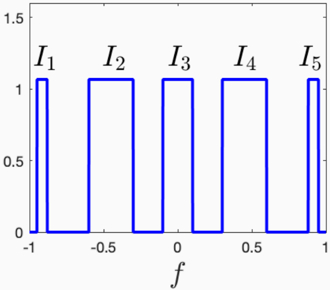
For these problems, we need to sample by an appropriately defined ridge leverage score:

$$\tau(t) = \max_{g} \frac{\frac{1}{T}\left|\mathcal{F}g(t)\right|^2}{\|\mathcal{F}g\|_2^2 + \epsilon\|g\|_2^2}$$



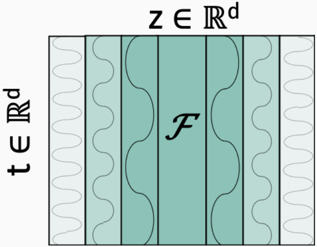
Sample and time efficient "spectrum blind" function fitting.





Ideally nearly matching the sample complexity of when the spectrum is known.

Everything in higher dimensions!



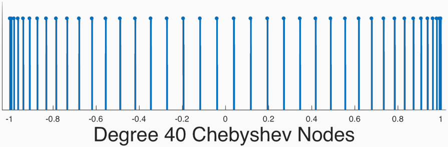
- **Sampling columns:** Random Fourier features methods for kernels (see our [ICML 2017] work).
- **Sampling rows:** Nystrom methods and active kernel regression.

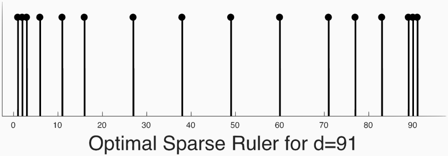
Understanding data distribution sampling.

"Relating Leverage Scores and Density using Regularized Christoffel Functions" – Pauwels, Bach, Vert 2018.

- In our problems we lost $O(q)/O(k)$ factors from data distribution sampling instead of leverage score sampling.
- For what other settings is this loss bounded? Is data distribution sampling okay under additional assumptions?

More explicit connections?





- How do deterministic methods in RandNLA (e.g. based on Batson. Spielman, Srivastava) compare with existing deterministic sampling schemes?
- How about methods for different norms?

THANK YOU!