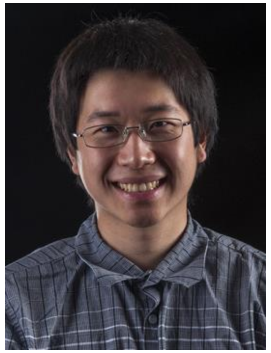
Fast Fourier Sparsity Testing

GRIGORY YAROSLAVTSEV

SAMSON ZHOU



The Alan Turing Institute

Carnegie Mellon University

Fourier Expansion

For $S \subseteq [n]$, the characteristic function $\chi_S(x): \{-1,+1\}^n \rightarrow \{-1,+1\}$ is defined as $\chi_S(x) = \prod_{i \in S} x_i$

The Fourier expansion of a function $f: \{-1,+1\}^n \rightarrow \mathbb{R}$ is the unique linear combination of multilinear polynomials:

$$f(x_1, \ldots, x_n) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x)$$

Then we have the Fourier coefficient $\hat{f}(S) = \langle f, \chi_S \rangle = E[f(x)\chi_S(x)]$

Fourier Sparsity

$$f(x_1, x_2, x_3) = Maj_3(x_1, x_2, x_3) = \frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{2}x_3 - \frac{1}{2}x_1 x_2 x_3$$

$$f(x_1, x_2, x_3, x_4, x_5, x_6) = 6 + 6x_1 x_2$$
$$= 3(x_1 + x_2)^2 - x_3^3 x_4 + 2x_3 x_4 x_5^2 - x_3 x_4^3 x_6^4$$

A function is s-sparse if it has at most s nonzero Fourier coefficients

Why Fourier Sparsity?

Fourier sparsity has applications in coding theory [GoldreichLevin89, AkaviaGoldwasserShafra03], learning theory [KushilevitzMansour93, LinialMansourNisan93], communication complexity [ShiZhang09]

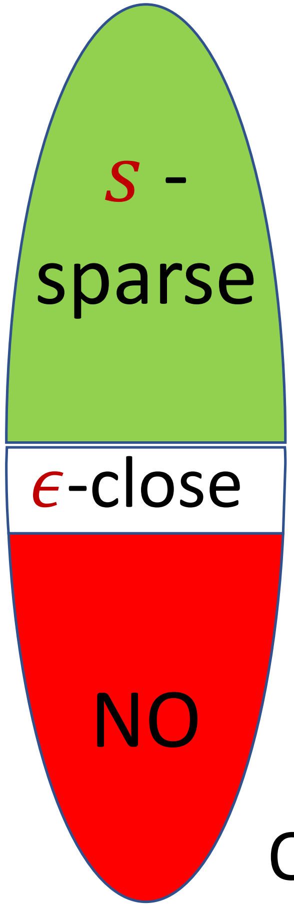
If a function is known to be s-sparse, more efficient algorithms can often be run, e.g. sparse Fourier transform [HIKP12]

Property Tester

Testing sparsity of Boolean functions under Hamming distance

- [GOSSW11]
  - Non-tolerant test
  - Complexity $O\left(s^{14}\log s+\frac{s^6}{\epsilon^2\log s}\right)$
  - Reduction to testing under ℓ_2
  - Lower bound $\Omega(\sqrt{s})$
- [WimmerYoshida13]
  - Tolerant test
  - Complexity $\mathrm{poly}\left(s,\frac{1}{\epsilon}\right)$

Our results give a tolerant test with almost quadratic improvement on [GOSSW11]



$\Rightarrow$ Accept with probability $\geq \frac{2}{3}$

$\Rightarrow$ Don't care

$\Rightarrow$ Reject with probability $\geq \frac{2}{3}$

Query model: $(x, f(x))$, for any $x \in \{-1,+1\}^n$

Our Contributions

Upper bound: Algorithm that makes $O\left(\frac{s}{\epsilon^4}\log\frac{1}{\epsilon}\right)$ non-adaptive queries to a (normalized) function f and approximates the ℓ_2^2 distance from f to the set of s-sparse functions. Translates to a tester.

Lower bound: Any algorithm that distinguishes whether f is s-sparse or at least $\frac{1}{3}$ - far from s-sparse in ℓ_2^2 distance requires $\Omega(\sqrt{s})$ queries

Random Hashing [FGKP09]

If a subspace H of $\{-1,+1\}^n$ is drawn randomly from subspaces of codimension d, then $\Pr[b \in a + H] = \frac{1}{2^d}$ for distinct $a, b \neq 0$

For an element $a \in H^\perp$, the projected function $f|_{a+H}(z) = E_{x \in H^\perp}[f(x+z)\chi_a(x)]$ for each $z \in \{-1,+1\}^n$

Poisson Summation Formula: $\hat{f}|_{a+H}(\alpha) = \hat{f}(\alpha)$ if $\alpha \in a + H$ and 0 otherwise

Define the total energy of $f|_{a+H}$ as $\sum_{\alpha \in a+H} \hat{f}(\alpha)^2 = \left\|\hat{f}|_{a+H}\right\|_2^2$

Testing s-Sparsity

 $= O(s^2) \Rightarrow$ 

$\sum_{\alpha \in a+H} \hat{f}(\alpha)^2 = E[\chi_a(z) f(x) f(x+z)]$, where $x \in \{-1,+1\}^n, z \in H^{\perp}$ [GOSSW11]

The set of queries $\{f(x+z)\}_{x \in H^{\perp}}$ can be used to compute $f|_{a+H}(z)$ for each of the cosets $a+H$ simultaneously

Algorithm

Draw subspace H of codimension $d = \log \frac{2s}{\epsilon^4}$ at random

Draw $O\left(\frac{s}{\epsilon^4}\right)$ pairs $(x, x + z)$, where $x \in \{-1, +1\}^n, z \in H^\perp$

Estimate the energy y_{a+H} for each $a \in H^\perp$

$$y_{a+H} \mathrel{+}= O\left(\frac{\epsilon^4}{s}\right) \chi_a(z) f(x) f(x+z)$$

Repeat $\ell = O\left(\log \frac{1}{\epsilon}\right)$ times, take the largest sum of the energies in s buckets

$$\xi = \max_{S \in H^\perp, |S| = s} \sum_{a \in S} \operatorname{median}\left(y_{a+H}^{(1)}, \dots, y_{a+H}^{(\ell)}\right)$$

Analysis

Top s coefficients may collide in a bucket

Noise from non top s coefficients

Hashing error: Let $y_1 \geq \cdots \geq y_s$ be the energies of the top s buckets and $f_1 \geq \cdots \geq f_s$ be the energies of the top s Fourier coefficients. Then with probability at least 15/16, the hashing error $\sum y_i - f_i \leq 5\epsilon^2$

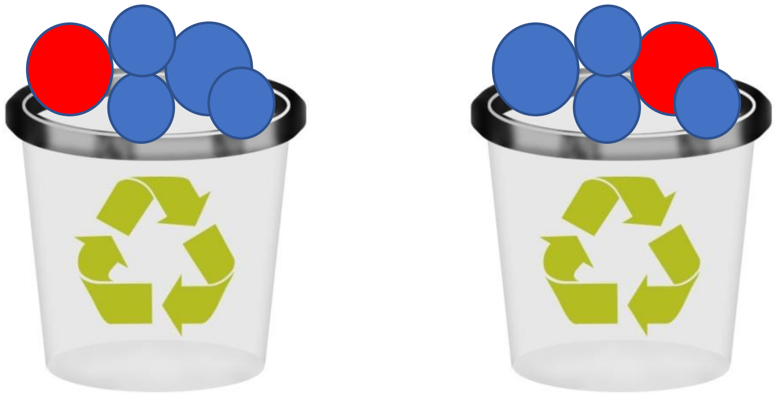
Estimation error: Let $y_1 \geq \cdots \geq y_s$ be the energies of the top s buckets and $\widehat{y}_i$ be the estimate of y_i. Then $E_H[\sum_{i=1}^{s} |\widehat{y}_i - y_i|^2] \leq \epsilon^2$

Putting it all together

Let S^* be the set of s buckets that maximize the estimated energies

Define h to be f with only the s Fourier coefficients that are the largest Fourier coefficient in each bucket, but their energies is the energy of the entire bucket

$\xi = \|h\|_2^2$



Putting it all together

Define f^* to be f with only the largest s Fourier coefficients

$$\left|\xi - \|f^*\|_2^2\right| = \left|\|h\|_2^2 - \|f^*\|_2^2\right| \leq 2\|f^* - h\|_2$$

Define g to be f with only the s Fourier coefficients that are the largest Fourier coefficient in each bucket

Then $\|f^* - h\|_2 \leq \underbrace{\|f^* - g\|_2}_{\text{Hashing error}} + \underbrace{\|g - h\|_2}_{\text{Estimation error}} = O(\epsilon)$

Future Work?

Improve upper or lower bounds

Extensions to other domains (line, hypergrid)

Other properties that can be tested in ℓ_2?

NSF


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
dziękuję
bedankt
спасибо
go raibh maith agat
mochchakkeram
sukriya
kop khun krap
terima kasih
dakujem
arigatô
grazie
takk
tapadh leat
dank je
감사합니다
ευχαριστώ
мерси
рахмат
Баярлалаа
enkosi
sağolun
hvala
mauruuru
blagodaram
kiitos
dankie
dhanyavad
faafetai lava
mersi
barka
welalin
tack
spas
paldies
grazzi
mahalo
misaotra
matondo
vinaka
хвала
asante
manana
obrigada
tenki
murakoze
chokrane
mamnun
дякую
djiere dieuf
tau
akun dankon
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
trugarez
shukriya
mercé
dhanyavadagalu
diolch
xièxie
rahmet
tanemirt
najis tuke
kam sah hamnida
ありがとう
mèsi
dèkuji
sobodi
bayarlalaa
gracie
köszönöm
nanni
nandri
didi madloba
rahmat
tau

Hashing Error

Hashing error: Let $y_1 \geq \cdots \geq y_s$ be the energies of the top s buckets and $f_1 \geq \cdots \geq f_s$ be the energies of the top s Fourier coefficients. Then with probability at least 15/16, the hashing error is at most $5\epsilon^2$ (Essentially just avoid collisions)

Define y_i^* as the largest Fourier coefficient hashing into bucket i

$$\sum y_i - f_i = \sum y_i - y_i^* + \sum y_i^* - f_i \leq \sum y_i - y_i^*$$

$E[\sum y_i - y_i^*] \leq \sqrt{\frac{2s}{2^d}}$, $\mathrm{Var}(\sum y_i - y_i^*) \leq \frac{2}{2^d}$ by Cauchy-Schwartz and Jensen, result holds by Chebyshev

Estimation Error

Estimation error: Let $y_1 \geq \cdots \geq y_s$ be the energies of the top s buckets and $\widehat{y_i}$ be the estimate of y_i. Then $E_H[\sum_{i=1}^{s} |\widehat{y_i} - y_i|^2] \leq \epsilon^2$

For a fixed $j \in [\ell]$, $E_H\left[\sum_{i=1}^{s} \left|y_{i,j} - y_i\right|^2\right] \leq \frac{\epsilon^4}{s}$ (number of samples)

By Markov and counting, $\Pr[|\widehat{y_i} - y_i|^2 \geq \eta] \leq \left(\frac{2e\epsilon^4}{s\eta}\right)^{\ell/2}$

Result follows from integrating probability density function and Jensen