Learning Set Functions that are Sparse in Non-Orthogonal Fourier Bases

ETH zürich

Chris Wendler, Andisheh Amrollahi, Bastian Seifert, Andreas Krause, Markus Püschel

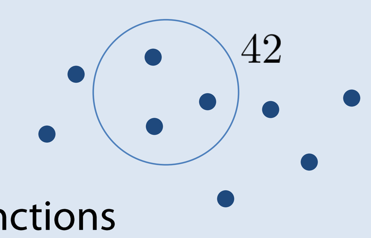
Department of Computer Science, ETH Zurich

Motivation

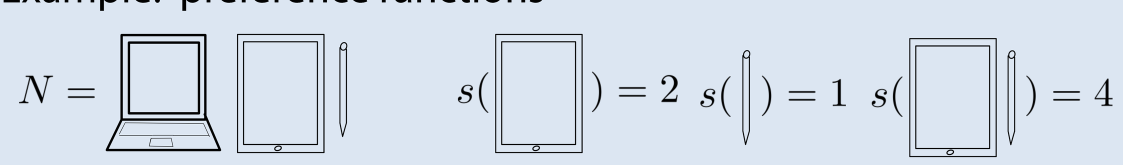
Set function Maps each subset to a real value

$N = \{x_1, \dots, x_n\}$

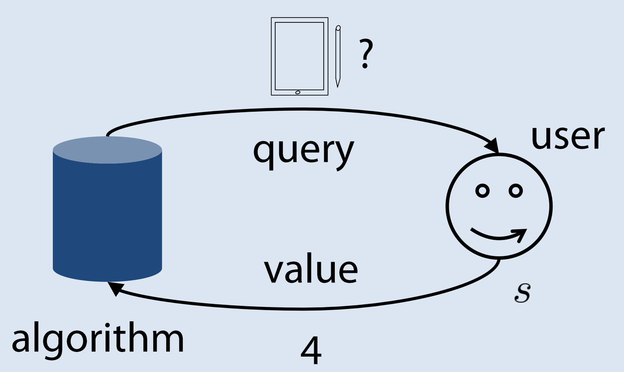
$s : 2^N \to \mathbb{R}$



Example: preference functions



Goal: Learn s from few queries

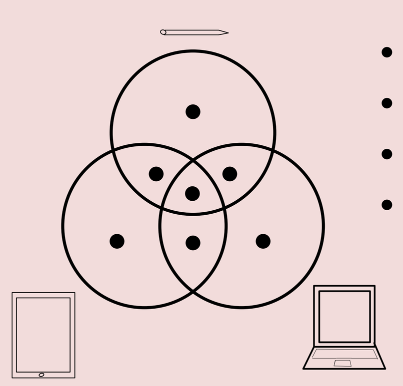


Challenge

Set functions are exponentially large (2^n)

Fourier-sparse set functions

New Non-Orthogonal Fourier Bases



- **new classes of learnable set functions**
- **new learning algorithms**
- **theoretical and experimental analysis**
- **interpretation:** new Fourier-sparsity captures complementarity and stubstitutability

Set Function Fourier Transforms

Derivation: Algebraic Signal Processing (ASP)

shift $\longrightarrow$ shift equivariant convolutional filters $\longrightarrow$ Fourier transform

Shift by $Q \subseteq N$ $\quad (T_Q s)(A) = s(A \cup Q)$

Convolutional filter

$$(h * s)(A) = \sum_{Q \subseteq N} h(Q) s(A \cup Q)$$

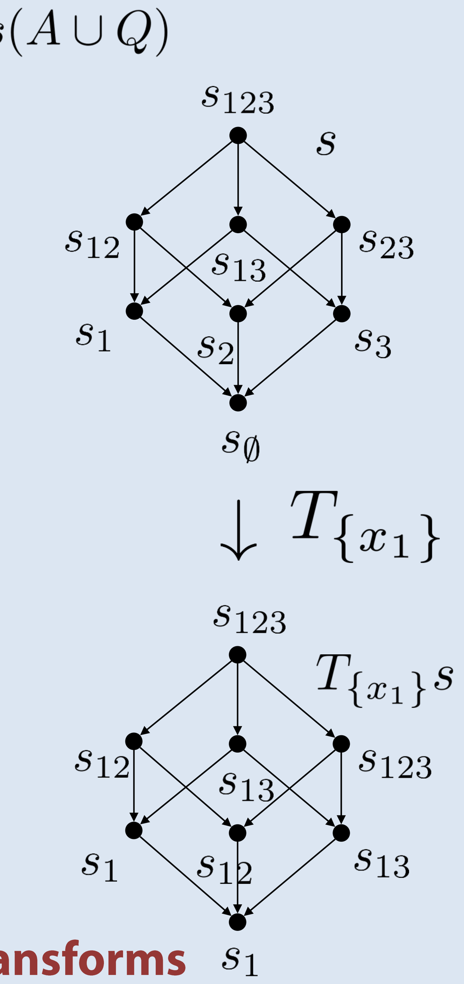
Equivariance

$$h * T_Q s = T_Q (h * s)$$

Fourier transform

$$\widehat{s}(B) = \sum_{A: A \cup B = N} (-1)^{|A \cap B|} s(A)$$

$$s(A) = \sum_{B: A \cap B = \emptyset} \widehat{s}(B)$$



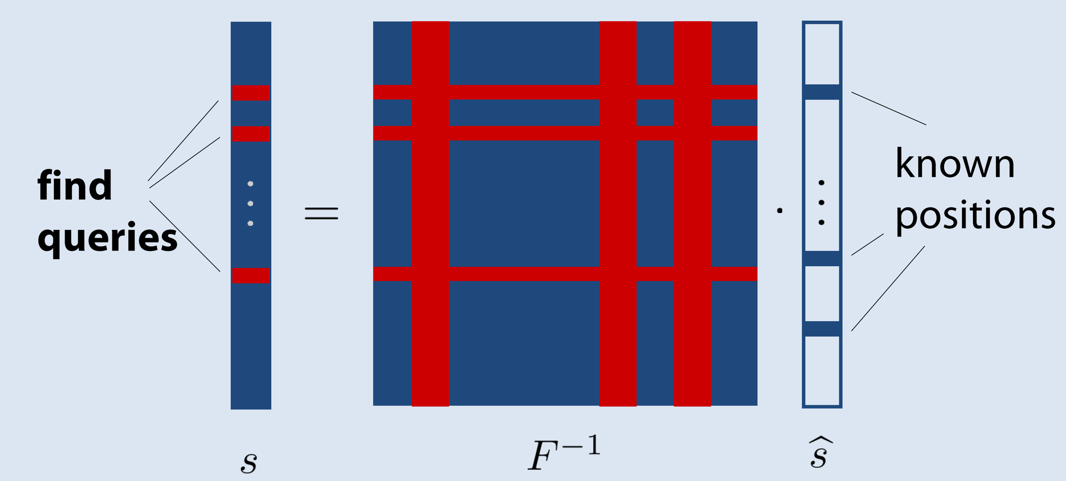
5 different shifts → 5 Fourier transforms

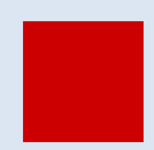
our paper: model 4 **prior:** Walsh Hadamard transform = model 5

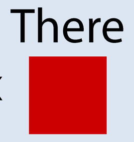
Known Support

Given oracle access to a *Fourier-sparse* set function s and its *support* $\mathcal{B} = \mathrm{supp}(\widehat{s}) = \{B_1, \dots, B_k\}$, learn s from few queries.

Idea



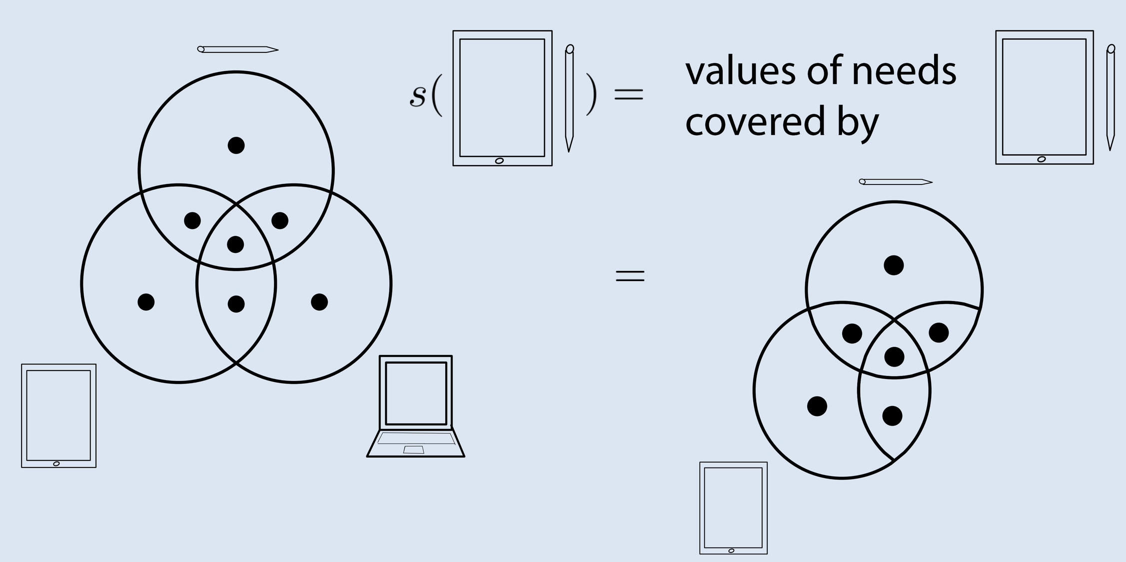
s.t. submatrix  **is invertible**

Theorem There always are queries s.t. the resulting submatrix  is triangular and invertible.

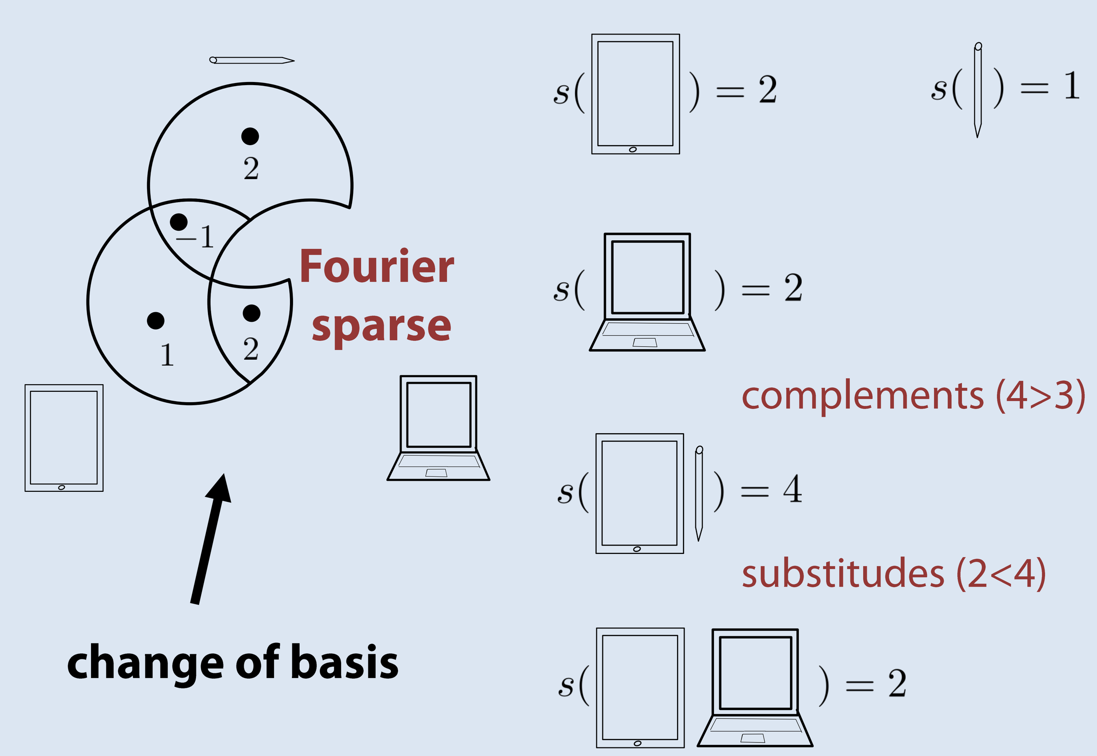
$\longrightarrow$ $O(k)$ queries, $O(k^2)$ operations

Fourier Sparsity

Set function = Venn diagram of needs



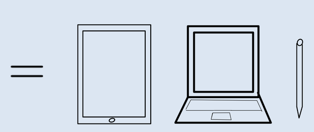
Fourier coefficients associate values to needs

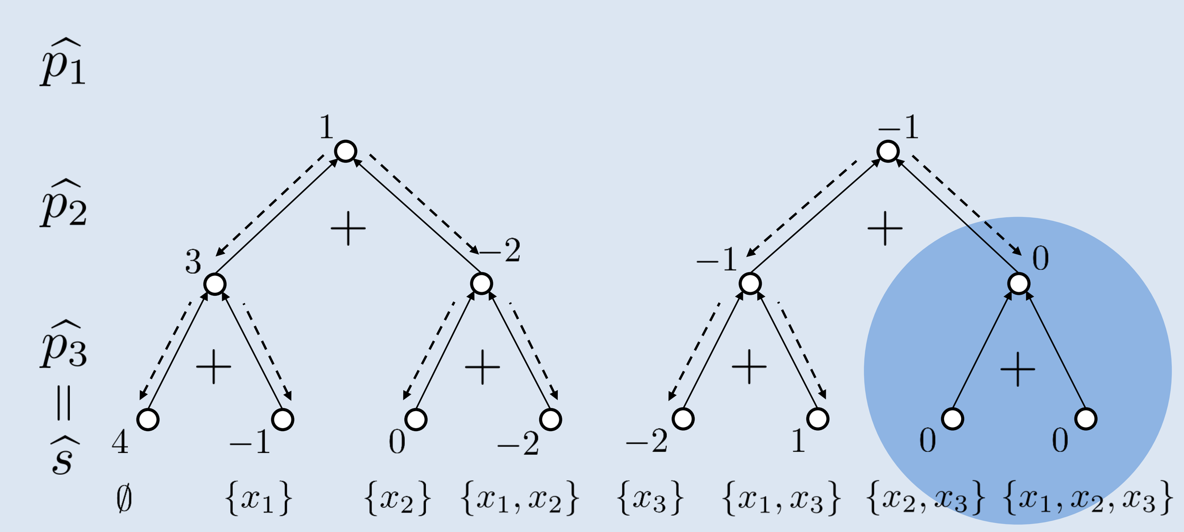


Unknown Support

Given oracle access to a *Fourier-sparse* set function s, learn s from few queries.

Idea Consider subproblems $p_1, p_2, \dots, p_n$:
$p_i(A) = s(A)$ for $A \subseteq \{y_1, \dots, y_i\} \subseteq N$

Consider $N = \{x_1, x_2, x_3\} =$ 



Observation $\mathrm{supp}(\widehat{p_{i+1}}) \overset{\text{SSFT}}{\longleftarrow} \mathrm{supp}(\widehat{p_i})$

Sparse Set Function Fourier Transform:

SSFT Algorithm avoids unnecessary computation

1. compute $\widehat{p_1}$ explicitely
2. compute $\widehat{p_i}$ from $\widehat{p_{i-1}}$ by solving the *known support* problem, for $i = 2, \dots, n$
3. return $\widehat{p_n} = \widehat{s}$

Queries

$O(nk)$ instead of 2^n

Operations **FFT**

$O(nk^2)$ instead of $n2^n$

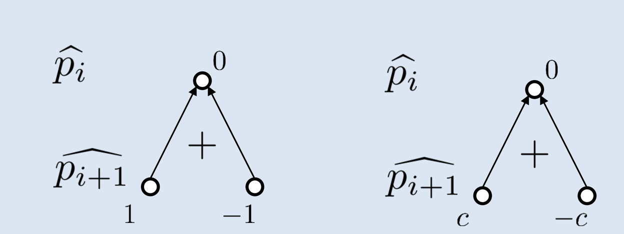
Refinement Algorithm

SSFT does not always work

Problem we want $\mathrm{supp}(\widehat{p_{i+1}}) \longleftarrow \mathrm{supp}(\widehat{p_i})$

we have $\mathrm{supp}(\widehat{p_{i+1}}) \dashleftarrow \mathrm{supp}(\widehat{p_i})$

Cancellations can occur

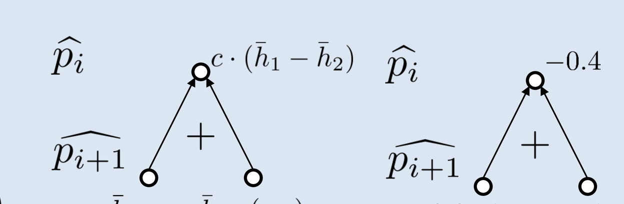


SSFT does not proccess children of 0's

Refinement Algorithm (SSFT+) use novel filtering techniques to ensure

Queries $O(n^2 k)$

Operations $O(n^2 k + nk^2)$



*Please find the details and analysis of SSFT(+) in our paper.

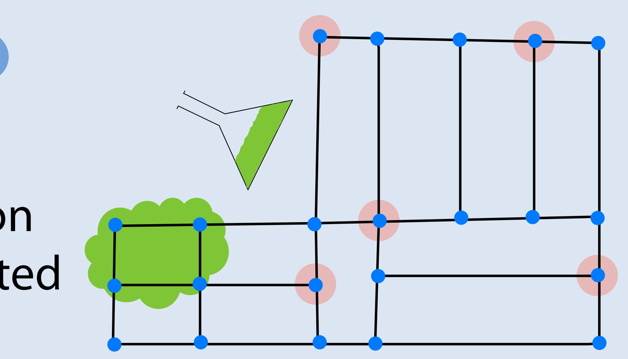
Experimental Evaluation

Facility Locations Functions

N = facility locations

s importance score

$s(A) = 50$ contamination events detected



Goal: learn importance score → find 5 best locations

Data from battle of water networks (Leskovec, 2007)

$\|N\|$		α	queries	time (s)	k	$\|\mathbf{p} - \mathbf{p}'\|/\|\mathbf{p}\|$
50	**SSFT**		**10K**	**2**	**648**	**0**
	R-WHT	1	*2,103K*	*361*	*1,380*	*0.001744*
		2	*4,192K*	*766*	*2,739*	*0.000847*
		4	*8,370K*	*1,499*	*5,054*	*0.000129*
		8	*16,742K*	*2,838*	*9,547*	*0.000108*
100	**SSFT**		**76K**	**24**	**2,308**	**0**
	R-WHT	1	16,544K	5,014	2,997	0.000546
		2	33,100K	10,265	6,466	0.000380
200	**SSFT**		**494K**	**451**	**7,038**	**0**
300	**SSFT**		**1,644K**	**2,368**	**16,979**	**0**
400	**SSFT**		**3,859K**	**7,654**	**28,121**	**0**
500	**SSFT**		**7,218K**	**17,693**	**38,471**	**0**

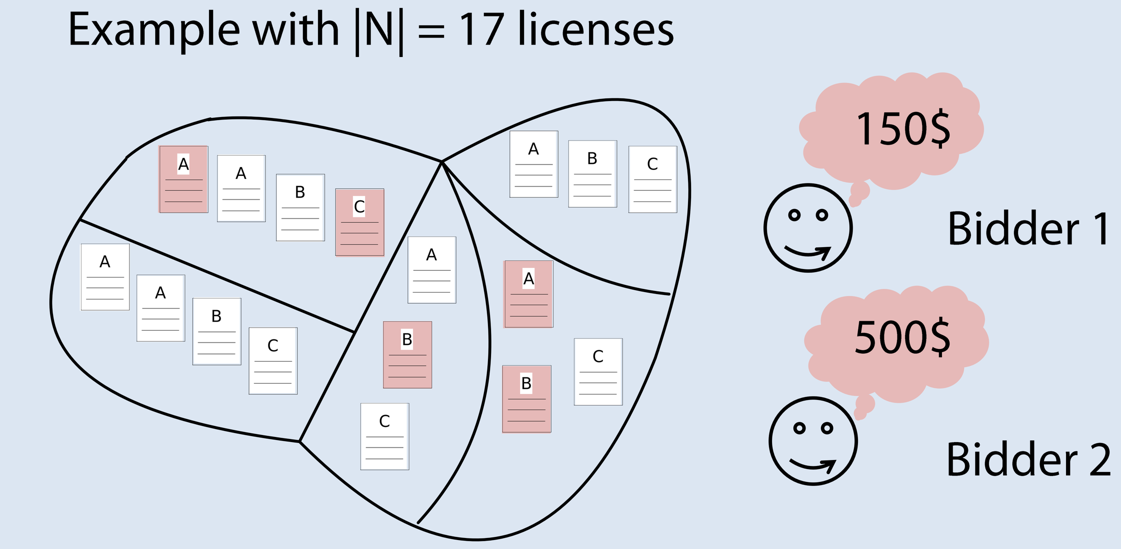
SSFT requires only 1,6 million queries to perfectly learn a 2^{300} dimensional set function.

Preference Elicitation in Combinatorial Spectrum Auctions

N = licenses of bands electromagnetic spectrum

Set functions = bidders

Example with |N| = 17 licenses



Goal: learn preferences → assign licenses to bidders

Multi region value model, |N| = 98, 3 bidder types (rows)

	number of queries (in thousands)		Fourier coefficients recovered		relative reconstruction error	
	SSFT+	**H-WHT**	**SSFT+**	**H-WHT**	**SSFT+**	**H-WHT**
L	229 ± 73	781 ± 0	303 ± 93	675 ± 189	0 ± 0	0 ± 0
R	646 ± 12	781 ± 0	813 ± 36	$1,779 \pm 0$	0 ± 0	0 ± 0
N	$3,305 \pm 1$	781 ± 0	$1,027 \pm 6$	$4,170 \pm 136$	0.01 ± 0.01	0.27 ± 0.21

SSFT+ perfectly learns the preferences of regional bidders using 656,000 queries and 813 Fourier coefficients.