Inner Product Similarity Join

Thomas D. Ahle, Rasmus Pagh, Ilya Razenshteyn, Francesco Silvestri

Main Theorem

Created by Samu Parra from the Noun Project

There are no subquadratic algorithms for any of the following problems:

1. Signed c-join for $c > 0$
2. Unsigned c-join for $c = e^{-o(\sqrt{\log n/\log\log n})}$
3. c-join of $P, Q \subseteq \{0, 1\}^d$ for $c = 1 - o(1)$
4. Unsigned c-join for $\log(cs)/\log(s) = 1 - o(1/\sqrt{\log n})$
5. c-join of $P, Q \subseteq \{0, 1\}^d$ for $\log(cs)/\log(s) = 1 - o(1/\log n)$

Problem definition

Exact IPS Join	c,s-Approx. Signed (c-join)	c-Approx. Unsigned
Given $P, Q \subseteq \mathbb{R}^n$ find all $x \in P$, $y \in Q$ st. $x^Ty > s$	Given $P, Q \subseteq \mathbb{R}^n$ find $x \in P$, $y \in Q$ st. $x^Ty > cs$ when it is guaranteed that a pair exists with inner product at least $s > cs > 0$	Given $P, Q \subseteq \mathbb{R}^n$ find $x \in P$, $y \in Q$ st. $\lvert x^Ty\rvert > cs$ when it is guaranteed that a pair exists with absolute inner product at least $s > cs$

Hardness in P "Understand hardness by reducing to well known hard problems"

- 3SUM: Given a set S of n integers, are there a,b,c 2 S with a+b+c = 0?
- Orthogonal vectors: Given a set S of n vectors in $\{0,1\}^d$, for $d = O(\log n)$ are there $x, y \in S$ with $x^Ty = 0$?
- All pairs shortest paths (APSP): given a weighted graph, find the distance between every two nodes.

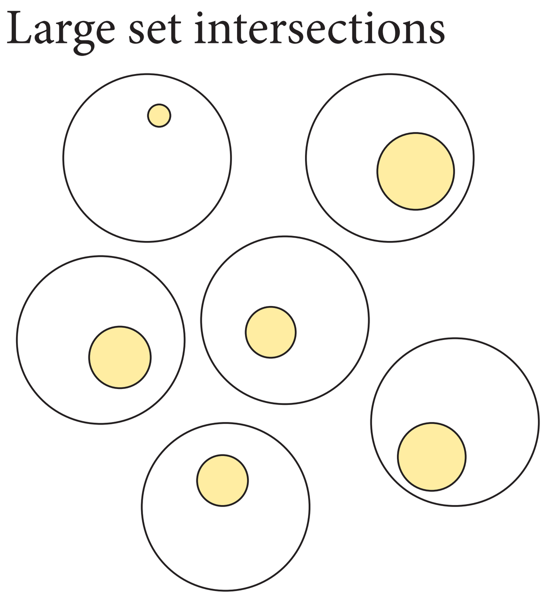
Orthogonal Vectors Conjecture

No algorithm solves OV in time $O(n^{2-\varepsilon})$ for $\varepsilon > 0$

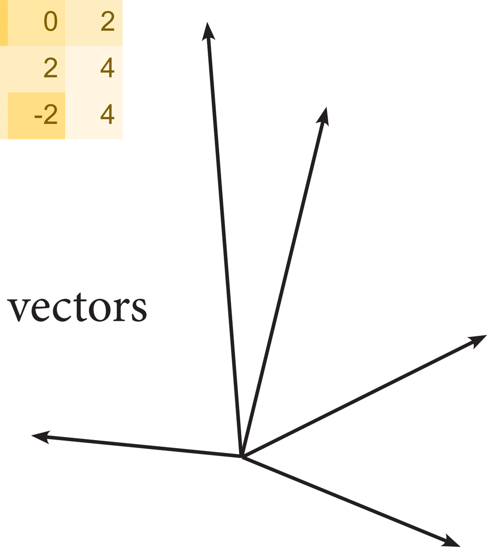
Large entries in matrix product

$$\begin{pmatrix} 1 & 1 & -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 1 & -1 & -1 & 1 \\ -1 & -1 & 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & 1 \\ -1 & -1 & 1 & -1 & -1 & -1 \\ -1 & -1 & 1 & 1 & 1 & -1 \end{pmatrix} = \begin{pmatrix} 0 & -4 & 0 & 2 & 2 & 4 \\ 2 & 6 & -2 & 0 & 0 & -2 \\ 0 & -4 & 0 & 2 & -2 & 0 \\ 2 & -2 & 2 & -4 & 0 & 2 \\ 0 & -4 & 0 & 2 & 2 & 4 \\ 4 & 0 & -4 & 2 & -2 & 4 \end{pmatrix}$$

Large set intersections



Highly correlated vectors



Polynomial evaluations

$4x^2 + 3xy - 2y^2 + 5$ $x, y = 3, 4$

$8x^2 + 6xy + y^2 - 1$ $x, y = -4, -5$

$x^2 - 20y^2 + 5$ $x, y = 0, 0$

Proof Idea

1 — We study embeddings $f, g : \{0,1\}^{d1} \to \mathbb{R}^{d2}$ so for every $x, y \in \{0,1\}^{d1}$
$\lvert f(x)^Tg(y)\rvert \le cs$, if $x^Ty \ge 1$
$\lvert f(x)^Tg(y)\rvert \ge s$, if $x^Ty = 0$
"Make an embedding"

2 — Make sure d2 (and the embedding time) is $n^{o(1)}$
"Apply the embedding"

3 — You've shown that unless OVP is false, there is no algorithm for unsigned (s,cs) join running in time $d^{O(1)}\, n^{2-\varepsilon}$
"Solve OVP"

Upper bounds - Exact IPS join

- Probabilistic Polynomials: $n^{2-1/O(c\log 2c)}$ for $P, Q \subseteq \{0,1\}^{c\log n}$ [Alman, Williams, FOCS 15]

Upper bounds - Approximate

- Data Dependent LSH: $n^{1+(1-s)/(1+s-2cs)}$ [Andoni, Razenshteyn STOC 15]
- Matrix multiplication: $n^{4/(3-\log s/\log cs)}$ [Karppa, Kaski, Kohonen, SODA 16]
- Sketching: For every $2 \le \kappa \le \infty$ there exists a distribution over $O(n^{1-2/\kappa}) \times n$ matrices Π such that for every $x \in R^n$ one has: $\Pr_\Pi[(1-c)\lVert x\rVert_\kappa \le \lVert \Pi x\rVert_\infty \le (1+c)\lVert x\rVert_\kappa] \ge 0.99$ [Andoni 10]

Simple embedding: showing c > 0 is impossible for the unsigned join

Take, coordinate wise

$f(0) := (1, -1, -1)$ $g(0) := (1, 1, -1)$

$f(1) := (1, 1, 1)$ $g(1) := (-1, -1, -1)$

Then

$f(1)^Tg(1) = -3$

$f(0)^Tg(1) = f(1)^Tg(0) = f(0)^Tg(0) = 1$

For vectors take

$f(x) = f(x_1) \dots f(x_n)\ 1^{d-4}$

$g(x) = g(x_1) \dots g(x_n)\ (-1)^{d-4}$

Such that

$f(x)^Tg(y) \le 0$,   if $x^Ty \ge 1$

$f(x)^Tg(y) = 4$,   if $x^Ty = 0$

For the other results, we use polynomial embeddings ~ [Valiant, FOCS 12]

{-1,1} Hardness by Chebyshev Polynomials

$T_n(x) = \cos(n\ \mathrm{acos}(x))$

$\lvert T_n(x)\rvert \le 1$ for $-1 \le x \le 1$

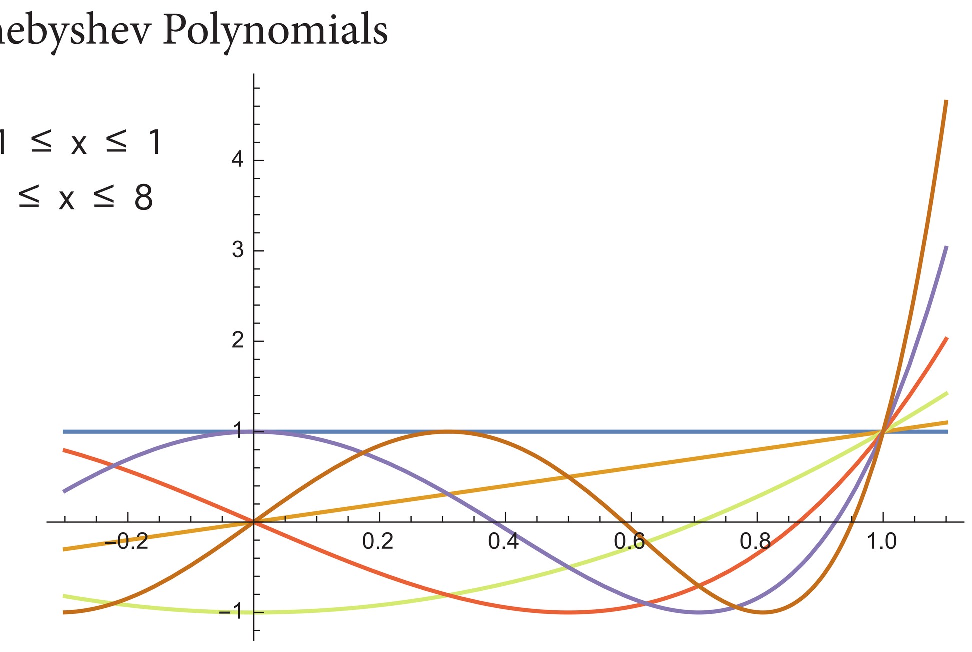
$T_n(x) \ge e^{n\sqrt{(x-1)}}$ for $1 \le x \le 8$

$T_0(x) = 1$

$T_1(x) = x$

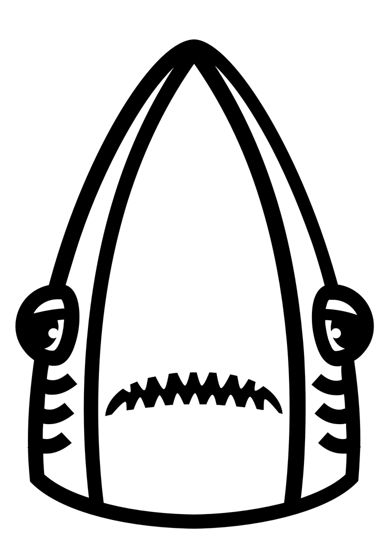
$T_2(x) = 2x^2 - 1$

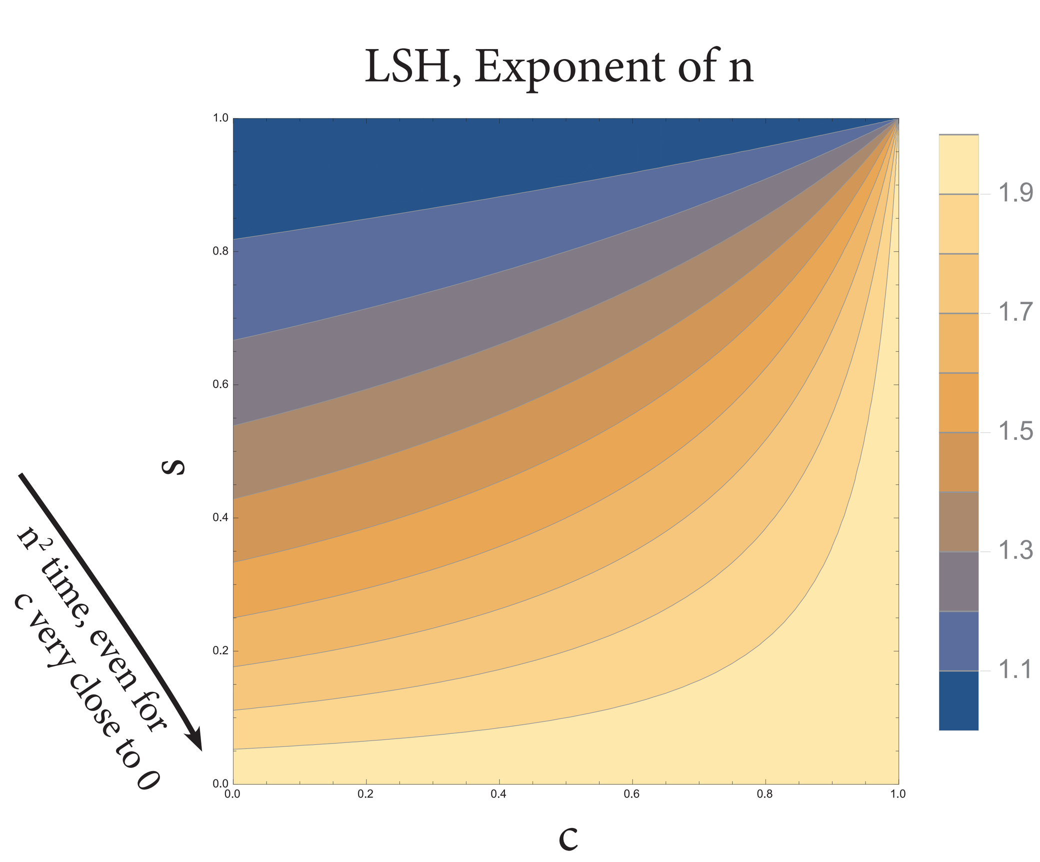
$T_3(x) = 4x^3 - 3x$

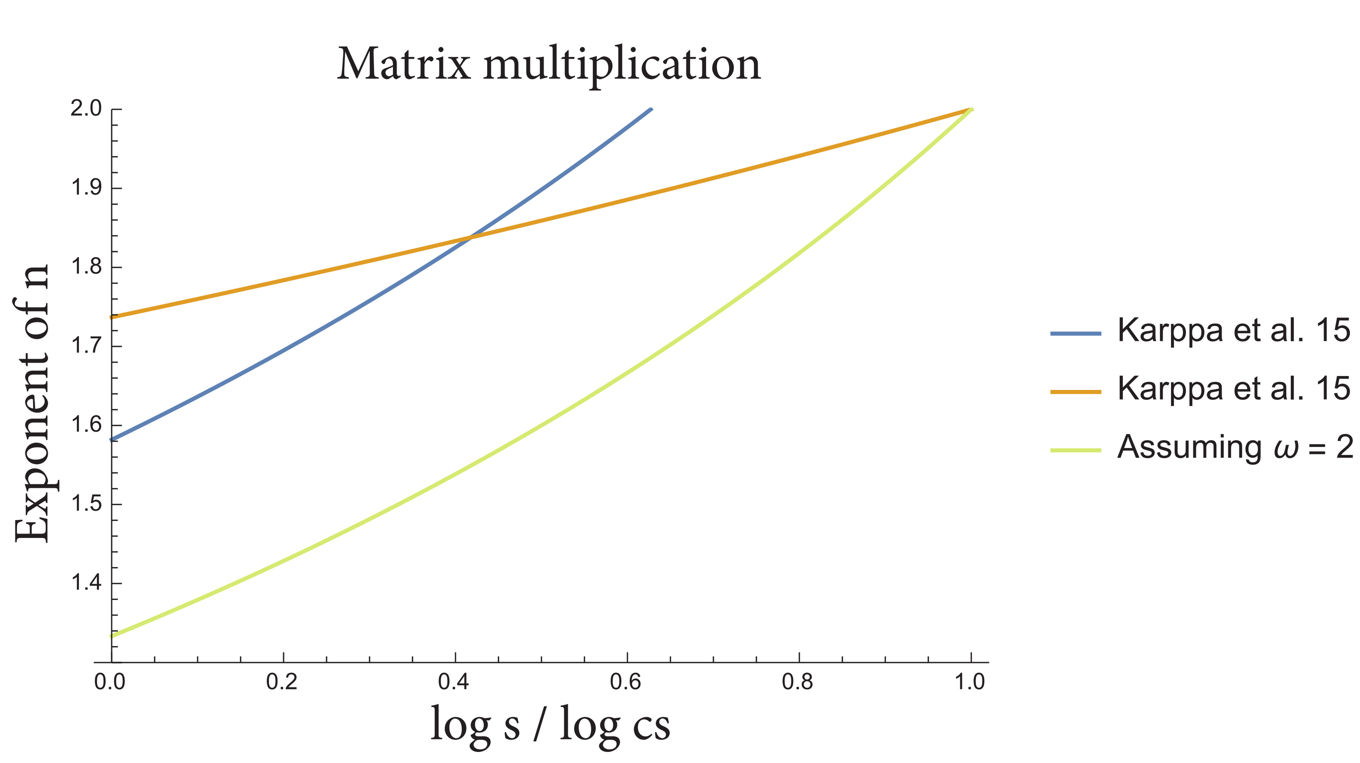


{0,1} Hardness by truncated product construction

$T(x) = (1-x_1y_1)\dots(1-x_ky_k) + (1-x_{k+1}y_{k+1})\dots(1-x_{2k}y_{2k}) + \dots$


Created by Jouk Fris from the Noun Project





Other results

- Hardness for datastructures, via reduction through assymetric OVP.
- Lower bound for LSH, "p1 - p2 = $O(\sqrt{s})$".
- Elimination of assymetry via error correcting codes.

Future work

- Get tighter bounds. Can we disprove subquadratic running time for logs/logcs = 1-eps?
- Prove or disprove separation between {-1,1} and {0,1} cases.
- Reconcile LSH and Matrix Multiplication methods.