Explicit Correlation Amplifiers for Finding Outlier Correlations in Deterministic Subquadratic Time[*][†]

Matti Karppa[1], Petteri Kaski[‡][2], Jukka Kohonen[3], and
Padraig Ó Catháin[4]

1 Helsinki Institute for Information Technology HIIT and Department of Computer Science, Aalto University, Helsinki, Finland
2 Helsinki Institute for Information Technology HIIT and Department of Computer Science, Aalto University, Helsinki, Finland
3 Helsinki Institute for Information Technology HIIT and Department of Computer Science, Aalto University, Helsinki, Finland
4 Helsinki Institute for Information Technology HIIT and Department of Computer Science, Aalto University, Helsinki, Finland

Abstract

We derandomize G. Valiant's [*J. ACM* 62 (2015) Art. 13] subquadratic-time algorithm for finding outlier correlations in binary data. Our derandomized algorithm gives deterministic subquadratic scaling essentially for the same parameter range as Valiant's randomized algorithm, but the precise constants we save over quadratic scaling are more modest. Our main technical tool for derandomization is an explicit family of *correlation amplifiers* built via a family of zigzag-product expanders in Reingold, Vadhan, and Wigderson [*Ann. of Math.* 155 (2002) 157–187]. We say that a function $f : \{-1,1\}^d \to \{-1,1\}^D$ is a *correlation amplifier* with threshold $0 \le \tau \le 1$, error $\gamma \ge 1$, and strength p an even positive integer if for all pairs of vectors $x, y \in \{-1,1\}^d$ it holds that (i) $|\langle x,y\rangle| < \tau d$ implies $|\langle f(x), f(y)\rangle| \le (\tau\gamma)^p D$; and (ii) $|\langle x,y\rangle| \ge \tau d$ implies $\left(\frac{\langle x,y\rangle}{\gamma d}\right)^p D \le \langle f(x), f(y)\rangle \le \left(\frac{\gamma\langle x,y\rangle}{d}\right)^p D$.

1998 ACM Subject Classification F.2.1 Numerical Algorithms and Problems

Keywords and phrases correlation, derandomization, outlier, similarity search, expander graph

Digital Object Identifier 10.4230/LIPIcs.ESA.2016.52

1 Introduction

Identifying weak correlations in data. We consider the task of finding outlier-correlated pairs from large collections of weakly correlated binary vectors in $\{-1,1\}^d$. In more precise terms, we are interested in the following computational problem.

▶ **Problem 1** (Outlier correlations). *Suppose we are given as input two sets $X, Y \subseteq \{-1,1\}^d$ with $|X| = |Y| = n$ and two thresholds, the* outlier *threshold $\rho > 0$ and the* background *threshold $\tau < \rho$. Our task is to output all* outlier *pairs $(x,y) \in X \times Y$ with $|\langle x,y\rangle| \ge \rho d$ subject to the assumption that at most q of the pairs $(x,y) \in X \times Y$ satisfy $|\langle x,y\rangle| > \tau d$.*

[*] We use the dagger-symbol ("†") to indicate a result whose proof for reasons of space has been relegated to the full version available at `http://arxiv.org/abs/1606.05608`.

[†] The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement 338077 "Theory and Practice of Advanced Search and Enumeration" (M.K., P.K., J.K.) and from the Academy of Finland, Grants 276031, 282938, 283262, and 283437 (P.Ó C.).

[‡] Work done in part while the second author was visiting the Simons Institute for the Theory of Computing.

24th Annual European Symposium on Algorithms (ESA 2016).
Editors: Piotr Sankowski and Christos Zaroliagis; Article No. 52; pp. 52:1–52:17
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

▶ Remark. This setting of binary vectors and (Pearson) correlation is directly motivated, among others, by the connection to Hamming distance. Indeed, for two vectors $x, y \in \{-1, 1\}^d$ we have $d - 2D_H(x, y) = \langle x, y\rangle$, where $D_H(x, y) = |\{u = 1, 2, \ldots, d : x(u) \neq y(u)\}|$ is the *Hamming distance* between x and y.

A naïve way to solve Problem 1 is to compute all the n^2 inner products $\langle x, y\rangle$ for $(x, y) \in X \times Y$ and filter out everything but the outliers. Our interest is in algorithms that scale *subquadratically* in n when both d and q are bounded from above by slowly growing functions of n. That is, we seek running times of the form $O(n^{2-\epsilon})$ for a constant $\epsilon > 0$. Furthermore, we seek to do this without *a priori* knowledge of q.

Running times of the form $O(n^{2-c\rho})$ for a constant $c > 0$ are immediately obtainable using techniques such as the seminal *locality-sensitive hashing* of Indyk and Motwani [17] and its variants[1]; however, such algorithms converge to quadratic running time in n unless ρ is bounded from below by a positive constant. Our interest is in algorithms that avoid such a "curse of weak outliers" and run in subquadratic time essentially *independently of the magnitude of ρ, provided that ρ is sufficiently separated from τ*. Such ability to identify weak outliers from large amounts of data is useful, among others, in machine learning from noisy data.

One strategy to circumvent the curse of weak outliers is to pursue the following intuition: (i) partition the input vectors into *buckets* of at most s vectors each, (ii) aggregate each bucket into a single vector by taking the vector sum, and (iii) compute the inner products between the $\lceil n/s\rceil \times \lceil n/s\rceil$ pairs of aggregate vectors. With sufficient separation between τ and ρ, at most q of these inner products between aggregates will be large, and every outlier pair is discoverable among the at most $s \times s$ input pairs that correspond to each large inner product of aggregates. Furthermore, a strategy of this form is oblivious to q until we actually start searching inside the buckets, which enables adjusting ρ and τ based on the number of large aggregate inner products.

Randomized amplification. Such bucketing strategies have been studied before with the help of randomization. In 2012, G. Valiant [33] presented a breakthrough algorithm that, before bucketing, replaces each input vector with a randomly subsampled[2] version of its p^{th} Kronecker power. Because of the tensor-power identity

$$\langle x^{\otimes p}, y^{\otimes p}\rangle = \langle x, y\rangle^p\,, \tag{1}$$

the ratio between outlier and background correlations gets *amplified* to essentially its p^{th} power, assuming that the sample is large enough so that sufficient concentration bounds hold with high probability. This amplification makes the outliers stand out from the background even after bucketing, which enables detection in subquadratic time using fast matrix multiplication.

A subset of the present authors [20] further improved on Valiant's algorithm by a modified sampling scheme that *simultaneously* amplifies and aggregates the input by further use of fast matrix multiplication. With this improvement, Problem 1 can be solved in subquadratic time if the logarithmic ratio $\log_\tau \rho = (\log \rho)/(\log \tau)$ is bounded from above by a constant less than 1. Also this improved algorithm relies on randomization.

[1] We postpone a more detailed discussion of related work and applications to the end of this section.

[2] Random sampling is used to reduce the dimension because the full d^p-dimensional Kronecker power is too large to be manipulated explicitly to yield subquadratic running times.

Explicit amplification. In this paper we seek *deterministic* subquadratic algorithms. As with the earlier randomized algorithms, we seek to map the d-dimensional input vectors to a higher dimension D so that inner products are sufficiently amplified in the process. Towards this end, we are interested in explicit functions $f : \{-1,1\}^d \to \{-1,1\}^D$ that approximate the tensor-power identity (1).

▶ **Definition 2** (Correlation amplifier). Let d, D and p be positive integers, with p even, and let $0 \le \tau \le 1$ and $\gamma \ge 1$. A function $f : \{-1,1\}^d \to \{-1,1\}^D$ is a *correlation amplifier* with parameters (d, D, p, τ, γ) if for all pairs of vectors $x, y \in \{-1,1\}^d$ we have

$$\text{if } \big|\langle x,y\rangle\big| < \tau d, \text{ then } \big|\langle f(x), f(y)\rangle\big| \le (\tau\gamma)^p D\,; \text{ and} \tag{2}$$

$$\text{if } \big|\langle x,y\rangle\big| \ge \tau d, \text{ then } \left(\tfrac{\langle x,y\rangle}{\gamma d}\right)^p D \le \langle f(x), f(y)\rangle \le \left(\tfrac{\gamma\langle x,y\rangle}{d}\right)^p D\,. \tag{3}$$

▶ Remark. A correlation amplifier f guarantees by (2) that correlations below τ in absolute value stay bounded; and by (3) that correlations at least τ in absolute value *become positive* and are governed by the two-sided approximation with multiplicative error $\gamma \ge 1$. In particular, (3) implies that correlations at least τ cannot mask outliers under bucketing because all such correlations get positive sign under amplification.

It is immediate that correlation amplifiers exist. For example, take $f(x) = x^{\otimes p}$, with p even, to obtain a correlation amplifier with $D = d^p$, $\tau = 0$, and $\gamma = 1$ by (1). For our present purposes, however, we seek correlation amplifiers with D substantially smaller than d^p. Furthermore, we seek constructions that are *explicit* in the strong[3] form that there exists a deterministic algorithm that computes any individual coordinate of $f(x)$ in time $\operatorname{poly}(\log D, p)$ by accessing $\operatorname{poly}(p)$ coordinates of a given $x \in \{-1,1\}^d$. In what follows explicitness always refers to this strong form.

Our results. The main result of this paper is that sufficiently powerful explicit amplifiers exist to find outlier correlations in deterministic subquadratic time.

▶ **Theorem 3** (Explicit amplifier family). *There exists an explicit correlation amplifier $f : \{-1,1\}^d \to \{-1,1\}^{2^K}$ with parameters $(d, 2^K, 2^\ell, \tau, \gamma)$ whenever $0 < \tau < 1$, $\gamma > 1$, and d, K, ℓ are positive integers with*

$$2^K \ge d\left(2^{10}\left(1-\gamma^{-1/2}\right)^{-1}\right)^{20\ell+1}\left(\frac{\gamma}{\tau}\right)^{60\cdot 2^\ell}. \tag{4}$$

As a corollary we obtain a deterministic algorithm for finding outlier correlations in subquadratic time using bucketing and fast matrix multiplication. Let us write α for the limiting exponent of rectangular integer matrix multiplication. That is, for all constants $\eta > 0$ there exists an algorithm that multiplies an $m \times \lfloor m^\alpha \rfloor$ integer matrix with an $\lfloor m^\alpha \rfloor \times m$ integer matrix in $O(m^{2+\eta})$ arithmetic operations. In particular, it is known that $0.3 < \alpha \le 1$ [22].

▶ **Theorem 4** (Deterministic subquadratic algorithm for outlier correlations). *For any constants $0 < \epsilon < 1$, $0 < \tau_{\max} < 1$, $0 < \delta < \alpha$, and $C > 60$, there exists a deterministic algorithm that solves a given instance of Problem 1 in time*

$$O\left(n^{2-\frac{0.99\epsilon(\alpha-\delta)}{4C+1}} + qn^{\delta+\frac{1.99\epsilon(\alpha-\delta)}{4C+1}}\right) \tag{5}$$

[3] In comparison, a weaker form of explicitness could require, for example, that there exists a deterministic algorithm that computes the entire vector $f(x)$ from a given x in time $D \cdot \operatorname{poly}(\log D, p)$.

assuming that the parameters n, d, ρ, τ satisfy the following three constraints

1. $d \le n^{\delta}$,
2. $n^{-\Theta(1)} \le \tau \le \tau_{\max}$, *and*
3. $\log_{\tau} \rho \le 1 - \epsilon$.

▶ Remark. Observe in particular that (5) is subquadratic regardless of the magnitude of ρ provided that the separation between ρ and τ via $\log_{\tau} \rho \le 1 - \epsilon$ holds.[4] The constants in (4) and (5) have not been optimized beyond our desired goal of obtaining deterministic subquadratic running time when d and q are bounded by slowly growing functions of n. In particular, (5) gives substantially worse subquadratic running times compared with the existing randomized strategies [20, 33]. The algorithm in Theorem 4 needs no *a priori* knowledge of q and is oblivious to q until it starts searching inside the buckets.

Overview and discussion of techniques. A straightforward application of the probabilistic method establishes that low-dimensional correlation amplifiers can be obtained by subsampling uniformly at random the dimensions of the tensor power $x^{\otimes p}$ as long as the sample size D is large enough.

▶ **Lemma 5** (Existence †). *There exists a correlation amplifier $f : \{-1,1\}^d \to \{-1,1\}^D$ with parameters (d, D, p, τ, γ) whenever $0 < \tau < 1$, $\gamma > 1$, and d, p, D are positive integers satisfying*

$$D \ge 3d\left(\gamma^p - 1\right)^{-2}\left(\frac{\gamma}{\tau}\right)^{2p} . \tag{6}$$

Thus, in essence our Theorem 3 amounts to *derandomizing* such a subsampling strategy by presenting an explicit sample that is, up to the error bounds (2) and (3), indistinguishable from the "perfect" amplifier $x \mapsto x^{\otimes p}$ under taking of inner products.

The construction underlying Theorem 3 amounts to an ℓ-fold composition of explicit *squaring* amplifiers ($p = 2$) with increasingly strong control on the error (γ) and the interval of amplification ($[\tau, 1]$) at each successive composition. Towards this end, we require a flexible explicit construction of squaring amplifiers with strong control on the error and the interval. We obtain such a construction from an explicit family of expander graphs (Lemma 9) obtainable from the explicit zigzag-product constructions of Reingold, Vadhan, and Wigderson [31]. In particular, the key to controlling the error and the interval is that the expander family gives *Ramanujan-like*[5] concentration $\lambda/\Delta \le 16\Delta^{-1/4}$ of the normalized second eigenvalue λ/Δ by increasing the degree Δ. In essence, since we are working with $\{-1,1\}$-valued vectors, by increasing the degree we can use the Expander Mixing Lemma (Lemma 8) and the Ramanujan-like concentration to control (Lemma 11) how well the restriction x^G to the edges of an expander graph G approximates the full tensor square $x^{\otimes 2}$ under taking of inner products.

Our construction has been motivated by the paradigm of *gradually increasing independence* [6, 11, 12, 18] in the design of pseudorandom generators. Indeed, we obtain the final amplifier gradually by successive squarings, taking care that the degree Δ_i of the expander

[4] The technical constraint $n^{-\Theta(1)} \le \tau$ only affects inputs where the dimension d grows essentially as a root function of n since $\tau \ge 1/d$. The constant subsumed by $\Theta(1)$ depends on the chosen constants $\epsilon, \tau_{\max}, \delta, C$ but not on the other parameters.

[5] Actual *Ramanujan graphs* (see [15, 23]) would give somewhat stronger concentration $\lambda/\Delta = O(\Delta^{-1/2})$ and hence improved constants in (4). However, we are not aware of a sufficiently fine-grained family of explicit Ramanujan graphs to comfortably support successive squaring.

that we apply in each squaring $i = 0, 1, \ldots, \ell - 1$ increases with a similar squaring schedule given by (10) and (12) to simultaneously control the error and the interval, and to bound the output dimension roughly by the square of the degree of the last expander in the sequence.[6] The analogy with pseudorandom generators can in fact be pushed somewhat further. Namely, a correlation amplifier can be roughly seen as a pseudorandom generator that by (3) seeks to fool a "truncated family of uniform combinatorial rectangles" with further control requested by (2) below the truncation threshold τ.[7] Our goal to obtain a small output dimension D roughly corresponds to optimizing the seed length of a pseudorandom generator.

While our explicit construction (4) does not reach the exact output dimension obtainable by Lemma 5, it should be observed that in our parameter range of interest (with $\gamma > 1$ a constant and $0 < \tau \leq \tau_{\max}$ for a constant $0 < \tau_{\max} < 1$), both (4) and (6) are of the form $D \geq d\tau^{-\Theta(p)}$; only the constants hidden by the asymptotic notation differ between the explicit and nonconstructive bounds. Moreover, using results of Alon [3] we show a *lower bound* on the output dimension D of any correlation amplifier: namely, that $D \geq d\tau^{-\Theta(p)}$ if $6p\tau^{-2}d^{-1}\log\frac{1}{\gamma\tau}$ is bounded from above by a constant strictly less than 1 (†). Thus, viewed as a pseudorandom generator with "seed length" $\log D$, Theorem 3 essentially does not admit improvement except possibly at the multiplicative constants.

Related work and applications. Problem 1 is a basic problem in data analysis and machine learning admitting many extensions, restrictions, and variants. A large body of work exists studying *approximate near neighbour search* via techniques such as locality-sensitive hashing (e.g. [4, 5, 17, 10, 26, 27]), with recent work aimed at derandomization (see Pagh [28] and Pham and Pagh [30]) and resource tradeoffs (see Kapralov [19]) in particular. However, these techniques enable subquadratic scaling in n only when ρ is bounded from below by a positive constant, whereas the algorithm in Theorem 4 remains subquadratic even in the case of weak outliers when ρ tends to zero with increasing n, as long as ρ and τ are separated. Ahle, Pagh, Razenshteyn, and Silvestri [1] show that subquadratic scaling in n is not possible for $\log_\tau \rho = 1 - o(1/\sqrt{\log n})$ unless both the Orthogonal Vectors Conjecture and the Strong Exponential Time Hypothesis [16] fail.

In small dimensions, Alman and Williams [2] present a randomized algorithm that finds *exact* Hamming-near neighbours in a batch-query setting analogous to Problem 1 in subquadratic time in n when the dimension is constrained to $d = O(\log n)$. Recently, Chan and Williams [7] show how to derandomize related algorithm designs, but the probabilistic polynomials for symmetric Boolean functions used in [2] to our knowledge have not yet been derandomized.

[6] The term "gradual" is of course not particularly descriptive since growth under successive squaring amounts to *doubly* exponential growth in the number of squarings. Yet such growth *can* be seen as gradual and controlled since we obtain strong amplification compared with the final output dimension precisely because the first $\ell - 1$ squarings "come for free" since $\Delta_0\Delta_1\cdots\Delta_{\ell-2}$ is (up to low-order multiplicative terms) no more than $\Delta_{\ell-1}^2$, essentially because we are taking the sum of powers of 2 in the exponent.

[7] To see the rough analogy, let $z \in \{-1,1\}^d$ be the Hadamard product of the vectors $x, y \in \{-1,1\}^d$ and observe that (3) seeks to approximate (with multiplicative error) the expectation of a uniform random entry in the d^p-length Kronecker power $z^{\otimes p}$ by instead taking the expectation over an explicit D-dimensional sample given by f. The Kronecker power $z^{\otimes p}$ is a uniform special case (with $z = z_1 = z_2 = \cdots = z_p$) of a "combinatorial rectangle" formed by a Kronecker product $z_1 \otimes z_2 \otimes \cdots \otimes z_p$, and truncation means that we only seek approximation in cases where $|\sum_{u=1}^{d} z(u)| \geq \tau d$, and accordingly want constructions that take this truncation into account—that is, we do not seek to fool all combinatorial rectangles and accordingly want stronger control on the dimension D (that is, the "seed length" $\log D$). For a review of the state of the art in pseudorandom generators we refer to Gopalan, Kane, and Meka [11] and Kothari and Meka [21].

One special case of Problem 1 is the problem of learning a weight 2 parity function in the presence of noise, or *the light bulb problem.*

▶ **Problem 6** (Light bulb problem, L. Valiant [34]). *Suppose we are given as input a parameter $0 < \rho < 1$ and a set of n vectors in $\{-1,1\}^d$ such that one planted pair of vectors has inner product at least ρd in absolute value, and all other $n-2$ vectors are chosen independently and uniformly at random. Our task is to find the planted pair among the n vectors.*

▶ Remark. From e.g. the Hoeffding bound (20) it follows that there exists a constant c such that when $d \geq c\rho^{-2}\log n$ the planted pair is with high probability (as n increases) the unique pair in the input with the maximum absolute correlation.

For a problem whose instances are drawn from a random ensemble, we say that an algorithm solves *almost all* instances of the problem if the probability of drawing an instance where the algorithm fails tends to zero as n increases.

Paturi, Rajasekaran, and Reif [29], Dubiner [8], and May and Ozerov [24] present randomized algorithms that can be used to solve almost all instances of the light bulb problem in subquadratic time if we assume that ρ is bounded from below by a positive constant; if ρ tends to zero these algorithms converge to quadratic running time in n.

G. Valiant [33] showed that a randomized algorithm can identify the planted correlation in subquadratic time on almost all inputs even when ρ tends to zero as n increases. As a corollary of Theorem 4, we can derandomize Valiant's design and still retain subquadratic running time (but with a worse constant) for almost all inputs, except for extremely weak planted correlations with $\rho \leq n^{-\Omega(1)}$ that our amplifier is not in general able to amplify with sufficiently low output dimension to enable an overall subquadratic running time.

▶ **Corollary 7** (Deterministic subquadratic algorithm for the light bulb problem). *For any constants $0 < \delta < \alpha$, $C > 60$, $0 < \rho_{\max} < 1$, and $\kappa > 1$, there exists a deterministic algorithm that solves almost all instances of Problem 6 in time*

$$O\left(n^{2-\frac{0.99(1-1/\kappa)(\alpha-\delta)}{4C+1}}\right)$$

assuming the parameters n, d, ρ satisfy the two constraints

1. $5\rho^{-2\kappa}\log n \leq d \leq n^{\delta}$ *and*
2. $n^{-\Theta(1)} \leq \rho \leq \rho_{\max}$.[8]

Corollary 7 extends to parity functions of larger (constant) weight (cf. [13, 20, 33]), however, we omit the details from this conference abstract. Algorithms for learning parity functions enable extensions to further classes of Boolean functions such as sparse juntas and DNFs (cf. [9, 25, 33]).

Conventions and notation. All vectors in this paper are integer-valued. For a vector $x \in \mathbb{Z}^d$ we denote the entry $u = 1, 2, \ldots, d$ of x by $x(u)$. For two vectors $x, y \in \mathbb{Z}^d$ we write $\langle x, y\rangle = \sum_{u=1}^{d} x(u)y(u)$ for the inner product of x and y. We write log for the logarithm with base 2 and ln for the logarithm with base exp(1).

[8] The constant hidden by the $\Theta(1)$ notation depends on the constants $\delta, \alpha, C, \rho_{\max}$ but not on the other parameters. For details consult the proof.

2 Explicit amplifiers by approximate squaring

This section proves Theorem 3. We start with preliminaries on expanders, show an approximate squaring identity using expander mixing, and then rely on repeated approximate squaring for our main construction. The proof is completed by some routine preprocessing.

Preliminaries on expansion and mixing. We work with undirected graphs, possibly with self-loops and multiple edges. A graph G is Δ-*regular* if every vertex is incident to exactly Δ edges, with each self-loop (if present) counting as one edge. Suppose that G is Δ-regular with vertex set V, and let L be a set of Δ labels such that the Δ edge-ends incident to each vertex have been labeled with unique labels from L. The *rotation map* $\mathrm{Rot}_G : V \times L \to V \times L$ is the bijection such that for all $u \in V$ and $i \in L$ we have $\mathrm{Rot}_G(u,i) = (v,j)$ if the edge incident to vertex u and labeled with i at u leads to the vertex v and has the label j at v.

For $S,T \subseteq V(G)$, let us write $E(S,T)$ for the set of edges of G with one end in S and the other end in T. Suppose that G has D vertices and let $\lambda_1, \lambda_2, \ldots, \lambda_D$ be the eigenvalues of the adjacency matrix of G with $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_D|$. Let us say that a graph G is a (D,Δ,λ)-*graph* if G has D vertices, G is Δ-regular, and $|\lambda_2| \leq \lambda$. For an excellent survey on expansion and expander graphs, we refer to Hoory, Linial, and Wigderson [15].

▶ **Lemma 8** (Expander mixing lemma, [15, Lemma 2.5]). *For all $S,T \subseteq V(G)$ we have*

$$\left| |E(S,T)| - \frac{\Delta |S||T|}{D} \right| \leq \lambda \sqrt{|S||T|}\,.$$

We work with the following family of graphs obtained from the zig-zag product of Reingold, Vadhan, and Wigderson [31]. In particular Lemma 9 gives us $\lambda/\Delta \leq 16\Delta^{-1/4}$, which will enable us to control relative inner products by increasing Δ.

▶ **Lemma 9.** *For all integers $t \geq 1$ and $b \geq 10$ there exists a $(2^{16bt}, 2^{4b}, 16 \cdot 2^{3b})$-graph whose rotation map can be evaluated in time* $\mathrm{poly}(b,t)$.[9]

Proof. See Appendix A. ◀

Main construction. The main objective of this section is to prove the following lemma, which we will then augment to Theorem 3 by routine preprocessing of the input dimension.

▶ **Lemma 10** (Repeated approximate squaring). *There exists an explicit correlation amplifier $\hat{f} : \{-1,1\}^{2^k} \to \{-1,1\}^{2^K}$ with parameters $(2^k, 2^K, 2^\ell, \tau_0, \gamma_0)$ whenever $0 < \tau_0 < 1$, $\gamma_0 > 1$, and k, K, ℓ are positive integers with*

$$2^K \geq 2^k \left(2^{10} \left(1 - \gamma_0^{-1}\right)^{-1} \right)^{20\ell} \left(\frac{\gamma_0}{\tau_0} \right)^{40 \cdot 2^\ell - 20} . \tag{7}$$

Approximate squaring via expanders. For a vector $x \in \{-1,1\}^D$, let us write $x^{\otimes 2} \in \{-1,1\}^{D^2}$ for the Kronecker product of x with itself. Our construction for correlation amplifiers will rely on approximating the *squaring identity*

$$\langle x^{\otimes 2}, y^{\otimes 2} \rangle = \langle x, y \rangle^2 \,,$$

[9] *Caveat.* Reingold, Vadhan, and Wigderson [31] work with eigenvalues of the *normalized* adjacency matrix (with $|\lambda_1| = 1$) whereas we follow Hoory, Linial, and Wigderson [15] and work with unnormalized adjacency matrices (with $|\lambda_1| = \Delta$) in the manuscript proper. Appendix A works with normalized adjacency matrices for compatibility with Reingold, Vadhan, and Wigderson [31].

for vectors in $\{-1,1\}^D$. In more precise terms, let G be a (D,Δ,λ)-graph and let $x^G \in \{-1,1\}^{\Delta D}$ be a vector that contains each coordinate $x(u)x(v)$ of $x^{\otimes 2}$ with $(u,v) \in V(G) \times V(G)$ exactly once for each edge of G that joins the vertex u to the vertex v. Equivalently, let $\mathrm{Rot}_G : V \times L \to V \times L$ be a rotation map for G, and define x^G for all $u \in V$ and all $i \in L$ by $x^G(u,i) = x(u)x(v)$ where $v \in V$ is given by $\mathrm{Rot}_G(u,i) = (v,j)$. In particular, x^G has exactly ΔD coordinates.

▶ **Lemma 11** (Approximate squaring). *For all $x,y \in \{-1,1\}^D$ we have*

$$\left| \langle x^G, y^G \rangle - \frac{\Delta}{D} \langle x^{\otimes 2}, y^{\otimes 2} \rangle \right| \le 2\lambda D\,.$$

Proof. Let $S = \{u \in V(G) : x(u) = y(u)\}$ and let us write $\bar{S} = V(G) \setminus S$. Since x,y are $\{-1,1\}$-valued, we have

$$\langle x^G, y^G \rangle = |E(S,S)| + |E(\bar{S},\bar{S})| - |E(S,\bar{S})| - |E(\bar{S},S)|\,.$$

Observing that

$$|S|^2 + |\bar{S}|^2 - |S||\bar{S}| - |\bar{S}||S| = \big(2|S| - D\big)^2 = \langle x,y \rangle^2 = \langle x^{\otimes 2}, y^{\otimes 2} \rangle$$

and applying Lemma 8 four times, we have

$$\left| \langle x^G, y^G \rangle - \frac{\Delta}{D} \langle x^{\otimes 2}, y^{\otimes 2} \rangle \right| \le \lambda\big(D + 2\sqrt{|S|(D - |S|)}\big) \le 2\lambda D\,.$$ ◀

The amplifier function. We now construct an amplifier function $\hat{f}$ that uses ℓ approximate squarings, $\ell \ge 1$, with the graphs drawn from the graph family in Lemma 9. Accordingly, we assume that all vectors have lengths that are positive integer powers of 2.

The input $x = \tilde{x}_0 \in \{-1,1\}^{d_0}$ to the amplifier has dimension $d_0 = 2^k$ for a positive integer k. For $i = 0,1,\ldots,\ell-1$, suppose we have the vector $\tilde{x}_i \in \{-1,1\}^{d_i}$. Let b_i be a positive integer whose value will be fixed later. Let t_i be the unique positive integer with

$$d_i \le D_i = 2^{16 b_i t_i} < 2^{16 b_i} d_i\,.$$

Note in particular that d_i divides D_i since d_i is a power of 2. Let G_i be a $(2^{16b_it_i}, 2^{4b_i}, 16\cdot 2^{3b_i})$-graph from Lemma 9. Take D_i/d_i copies of $\tilde{x}_i$ to obtain the vector $x_i \in \{-1,1\}^{D_i}$. Let $\tilde{x}_{i+1} = x_i^{G_i} \in \{-1,1\}^{d_{i+1}}$ with $d_{i+1} = \Delta_i D_i$ and $\Delta_i = 2^{4b_i}$. The amplifier outputs $\hat{f}(x) = \tilde{x}_\ell$ with $\tilde{x}_\ell \in \{-1,1\}^{d_\ell}$.

Since the graph family in Lemma 9 admits rotation maps that can be computed in time $\mathrm{poly}(b,t)$, we observe that $\hat{f}$ is explicit. Indeed, from the construction it is immediate that to compute any single coordinate of $\hat{f}(x)$ it suffices to (i) perform in total $2^{\ell-1-i}$ evaluations of the rotation map of the graph G_i for each $i = 0,1,\ldots,\ell-1$, and (ii) access at most 2^ℓ coordinates of x. Since $b_it_i = O(\log d_\ell)$ for all $i = 0,1,\ldots,\ell-1$, we have that we can compute any coordinate of $\hat{f}(x)$ in time $\mathrm{poly}(\log d_\ell, 2^\ell)$ and accessing at most 2^ℓ coordinates of x.

Parameterization and analysis. Fix $\tau_0 > 0$ and $\gamma_0 > 1$. To parameterize the amplifier (that is, it remains to fix the values b_i), let us track a pair of vectors as it proceeds through the ℓ approximate squarings for $i = 0,1,\ldots,\ell-1$.

We start by observing that copying preserves *relative* inner products. That is, for any pair of vectors $\tilde{x}_i, \tilde{y}_i \in \{-1,1\}^{d_i}$ we have $\langle \tilde{x}_i, \tilde{y}_i \rangle = \nu_i d_i$ if and only if $\langle x_i, y_i \rangle = \nu_i D_i$ for $0 \le \nu_i \le 1$.

An easy manipulation of Lemma 11 using the parameters in Lemma 9 gives us additive control over an approximate squaring via

$$\nu_i^2 - 32\Delta_i^{-1/4} \le \nu_{i+1} \le \nu_i^2 + 32\Delta_i^{-1/4} . \tag{8}$$

For all inner products that are in absolute value above a threshold, we want to turn this additive control into multiplicative control via

$$\nu_i^2\gamma_0^{-1} \le \nu_{i+1} \le \nu_i^2\gamma_0 . \tag{9}$$

Let us insist this multiplicative control holds whenever $|\nu_i| \ge \tau_i$ for the threshold parameter τ_i defined for all $i = 0, 1, \ldots, \ell - 1$ by

$$\tau_{i+1} = \gamma_0^{-1}\tau_i^2 . \tag{10}$$

Enforcing (9) via (8) at the threshold, let us assume that

$$\tau_i^2\gamma_0^{-1} \le \tau_i^2 - 32\Delta_i^{-1/4} . \tag{11}$$

The next lemma confirms that assuming (11) gives two-sided control of inner products which is retained to the next approximate squaring. The following lemma shows that small inner products remain small.

▶ **Lemma 12** (†). *If $\tau_i \le |\nu_i|$, then $\nu_i^2\gamma_0^{-1} \le \nu_{i+1} \le \nu_i^2\gamma_0$ and $\tau_{i+1} \le \nu_{i+1}$.*

▶ **Lemma 13** (†). *If $|\nu_i| < \tau_i$, then $|\nu_{i+1}| \le \tau_i^2\gamma_0$.*

Let us now make sure that (11) holds. Solving for Δ_i in (11), we have

$$\Delta_i \ge \left(32(1-\gamma_0^{-1})^{-1}\tau_i^{-2}\right)^4 . \tag{12}$$

In particular, we can make sure that (12) and hence (11) holds by simply choosing a large enough Δ_i (that is, a large enough b_i).

Before proceeding with the precise choice of b_i for $i = 0, 1, \ldots, \ell-1$, let us analyze the input–output relationship of the amplifier $\hat{f}$ using Lemma 12 and Lemma 13. Let $x, y \in \{-1, 1\}^{d_0}$ be two vectors given as input with $\langle x, y\rangle = \nu_0 d_0$. The outputs $\hat{f}(x), \hat{f}(y) \in \{-1, 1\}^{d_\ell}$ then satisfy $\langle \hat{f}(x), \hat{f}(y)\rangle = \nu_\ell d_\ell$, where the following two lemmas control ν_ℓ via ν_0.

▶ **Lemma 14** (†). *If $|\nu_0| \ge \tau_0$, then $\nu_0^{2^\ell}\gamma_0^{-2^\ell+1} \le \nu_\ell \le \nu_0^{2^\ell}\gamma_0^{2^\ell-1}$.*

▶ **Lemma 15** (†). *If $|\nu_0| < \tau_0$, then $|\nu_\ell| \le \tau_0^{2^\ell}\gamma_0^{2^\ell-1}$.*

Since $\gamma_0 > 1$, from Lemma 14 and Lemma 15 it now follows that $\hat{f}$ meets the required amplification constraints (2) and (3) with $p = 2^\ell$, $\tau = \tau_0$, and $\gamma = \gamma_0$.

Let us now complete the parameterization and derive an upper bound for d_ℓ. For each $i = 0, 1, \ldots, \ell-1$, take b_i to be the smallest nonnegative integer so that $b_i \ge 10$ and $\Delta_i = 2^{4b_i}$ satisfies (12). Since $D_i \le 2^{16b_i}d_i = \Delta_i^4 d_i$, we have $d_{i+1} = \Delta_i D_i \le \Delta_i^5 d_i$, and hence

$$d_\ell \le (\Delta_{\ell-1}\Delta_{\ell-2}\cdots\Delta_0)^5 d_0 .$$

Recall that $d_0 = 2^k$. From (12) we have that

$$\Delta_i = 2^{4b_i} \le \max\left(2^{40}, 2^4\left(32(1-\gamma_0^{-1})^{-1}\tau_i^{-2}\right)^4\right) \le \left(2^{10}(1-\gamma_0^{-1})^{-1}\tau_i^{-2}\right)^4 .$$

Since $\tau_i = \tau_0^{2^i}\gamma_0^{-2^i+1}$ by (10), it follows that

$$d_\ell \le 2^k\left(2^{10}\left(1-\gamma_0^{-1}\right)^{-1}\right)^{20\ell}\left(\frac{\gamma_0}{\tau_0}\right)^{20(2^{\ell+1}-1)} .$$

Repeatedly taking two copies of the output as necessary, for all 2^K with $2^K \ge d_\ell$ we obtain a correlation amplifier with parameters $(2^k, 2^K, 2^\ell, \tau_0, \gamma_0)$. This completes the proof of Lemma 10. ◀

Copy-and-truncate preprocessing of the input dimension. We still want to remove the assumption from Lemma 10 that the input dimension is a positive integer power of 2. The following copy-and-truncate preprocessing will be sufficient towards this end.

Let $x \in \{-1,1\}^d$ and let k be a positive integer. Define the vector $\hat{x} \in \{-1,1\}^{2^k}$ by concatenating $\lceil 2^k/d \rceil$ copies of x one after another, and truncating the result to the 2^k first coordinates to obtain $\hat{x}$.

Let us study how the map $x \mapsto \hat{x}$ operates on a pair of vectors $x, y \in \{-1,1\}^d$. For notational compactness, let us work with relative inner products $\nu, \hat{\nu}$ with $\langle x, y\rangle = \nu d$ and $\langle \hat{x}, \hat{y}\rangle = \hat{\nu} 2^k$.

▶ **Lemma 16** (†). *For any $0 < \tau_0 < 1$, $\gamma_0 > 1$, and $2^k \geq 2d\tau_0^{-1}(1-\gamma_0^{-1})^{-1}$ we have that*
1. $|\nu| < \tau_0$ *implies* $|\hat{\nu}| \leq \gamma_0\tau_0$*,*
2. $|\nu| \geq \tau_0$ *implies* $\gamma_0^{-1}\nu \leq |\hat{\nu}| \leq \gamma_0\nu$*.*

Completing the proof of Theorem 3. Let d, K, ℓ, τ, γ be parameters meeting the constraints in Theorem 3, in particular the constraint (4). To construct a required amplifier f, we preprocess each input vector x with copy-and-truncate, obtaining a vector $\hat{x}$ of length 2^k. We then then apply an amplifier $\hat{f} : \{-1,1\}^{2^k} \to \{-1,1\}^{2^K}$ given by Lemma 10. In symbols, we define $f : \{-1,1\}^d \to \{-1,1\}^{2^K}$ for all $x \in \{-1,1\}^d$ by $f(x) = \hat{f}(\hat{x})$. It is immediate from Lemma 10 and Lemma 16 that the resulting composition is explicit.

We begin by relating the given parameters of Theorem 3 to those of Lemma 10. Take $\gamma_0 = \gamma^{1/2}$, $\tau_0 = \tau\gamma^{-1}$, and select the minimal value of k so that the constraint in Lemma 16 is satisfied; that is 2^k is constrained as follows,

$$2d(1-\gamma^{-1/2})^{-1}\gamma\tau^{-1} \leq 2^k < 4d(1-\gamma^{-1/2})^{-1}\gamma\tau^{-1}\,.$$

Substituting this upper bound into the bound of Lemma 10, we get a lower bound for 2^K,

$$2^K \geq 2^{-8} d \left(2^{-10}(1-\gamma^{-1/2})^{-1}\right)^{20\ell+1} \frac{\gamma}{\tau} \left(\frac{\gamma^{60}}{\tau^{40}}\right)^{2^\ell} \frac{\tau^{20}}{\gamma^{30}}\,. \tag{13}$$

Observe that an integer 2^K satisfying (4) also satisfies (13). We have not attempted to optimise our construction, and prefer the the statement of Theorem 3 as it is reasonably clean and is sufficient to prove Theorem 4.

Let us study how the map $x \mapsto f(x)$ operates on a pair of vectors $x, y \in \{-1,1\}^d$. For notational compactness, again we work with relative inner products $\nu, \hat{\nu}, \phi$ with $\langle x, y\rangle = \nu d$, $\langle \hat{x}, \hat{y}\rangle = \hat{\nu}2^k$, and $\langle f(x), f(y)\rangle = \phi 2^K$. Observe that in the notation of the proof of Lemma 10, we have $\hat{\nu} = \nu_0$ and $\phi = \nu_\ell$.

▶ **Lemma 17** (†). *If $|\nu| < \tau$ then $|\phi| \leq (\gamma\tau)^{2^\ell}$.*

▶ **Lemma 18** (†). *If $|\nu| \geq \tau$ then $(\nu\gamma^{-1})^{2^\ell} \leq \phi \leq (\nu\gamma)^{2^\ell}$.*

Now, f satisfies (2) and (3) with $p = 2^\ell$ by Lemmas 17 and 18 respectively.

This completes the proof of Theorem 3. ◀

3 A deterministic algorithm for outlier correlations

This section proves Theorem 4. We start by describing the algorithm, then parameterize it and establish its correctness, and finally proceed to analyze the running time.

The algorithm. Fix the constants $\epsilon, \tau_{\max}, \delta, C$ as in Theorem 4. Based on these constants, fix the constants $0 < \sigma < 1$ and $\gamma > 1$. (We fix the precise values of σ and γ later during the analysis of the algorithm, and stress that σ, γ do not depend on the given input.)

Suppose we are given as input the parameters $0 < \tau < \rho < 1$ and $X, Y \subseteq \{-1,1\}^d$ with $|X| = |Y| = n$ so that the requirements in Theorem 4 hold. We work with a correlation amplifier $f : \{-1,1\}^d \to \{-1,1\}^D$ with parameters (d, D, p, τ, γ). (We fix the precise values of the parameters p and D later during the analysis of the algorithm so that f originates from Theorem 3.)

The algorithm proceeds as follows. First, apply f to each vector in X and Y to obtain the sets X_f and Y_f. Let $s = \lfloor n^\sigma \rfloor$. Second, partition the n vectors in both X_f and Y_f into $\lceil n/s \rceil$ buckets of size at most s each, and take the vector sum of the vectors in each bucket to obtain the sets $\tilde{X}_f, \tilde{Y}_f \subseteq \{-s, -s+1, \ldots, s-1, s\}^D$ with $|\tilde{X}_f|, |\tilde{Y}_f| \leq \lceil n/s \rceil$. Third, using fast rectangular matrix multiplication on $\tilde{X}_f$ and $\tilde{Y}_f$, compute the matrix Z whose entries are the inner products $\langle \tilde{x}, \tilde{y} \rangle$ for all $\tilde{x} \in \tilde{X}_f$ and all $\tilde{y} \in \tilde{Y}_f$. Fourth, iterate over the entries of Z, and whenever the *detection inequality*

$$\langle \tilde{x}, \tilde{y} \rangle > n^{2\sigma} (\tau\gamma)^p \tag{14}$$

holds, brute-force search for outliers among the at most $s \times s$ inner products in the corresponding pair of buckets. Output any outliers found.

Parameterization and correctness. Let us now parameterize the algorithm and establish its correctness. Since $\gamma > 1$ is a constant and assuming that p is large enough, by Theorem 3 we can select D to be the integer power of 2 with

$$\frac{1}{2} d \left(\frac{\gamma}{\tau}\right)^{Cp} < D \leq d \left(\frac{\gamma}{\tau}\right)^{Cp} .$$

Recall that we write α for the exponent of rectangular matrix multiplication. To apply fast rectangular matrix multiplication in the third step of the algorithm, we want

$$D \leq 2 \left(\frac{n}{s}\right)^{\alpha} , \tag{15}$$

so recalling that $d \leq n^\delta$ and $n^\sigma - 1 < s$, it suffices to require that

$$\left(\frac{\gamma}{\tau}\right)^{Cp} \leq n^{(1-\sigma)\alpha - \delta} .$$

Let us assume for the time being that $(1-\sigma)\alpha - \delta > 0$. (We will justify this assumption later when we choose a value for σ.) Let p be the unique positive-integer power of 2 such that

$$\frac{((1-\sigma)\alpha - \delta)\log n}{2C \log \frac{\gamma}{\tau}} < p \leq \frac{((1-\sigma)\alpha - \delta)\log n}{C \log \frac{\gamma}{\tau}} . \tag{16}$$

Observe that p exists and is positive for all large enough n since $\gamma > 1$ is a constant and $n^{-\Theta(1)} \leq \tau$ by our assumption.[10] By the detection inequality (14), we require each entry of Z to have value strictly greater than $n^{2\sigma}(\tau\gamma)^p$ if among the correspoding at most $s \times s$

[10] In particular, since $\sigma, \delta, \alpha, C, \gamma$ are constants, we can choose the constant hidden by the $\Theta(1)$ so that $1 \leq (((1-\sigma)\alpha - \delta)\log n)/(2C \log \frac{\gamma}{\tau})$.

inner products between the two buckets there is at least one inner product with absolute value at least ρd. Furthermore, we want (14) to hold only if among the at most $s \times s$ inner products between the two buckets there is at least one inner product with absolute value strictly greater than τd. Since f satisfies (2) and (3), and recalling that $s \le n^{\sigma}$, it suffices to require that

$$s^2(\tau\gamma)^p \le n^{2\sigma}(\tau\gamma)^p < (\rho\gamma^{-1})^p - n^{2\sigma}(\tau\gamma)^p\,. \tag{17}$$

Rearranging the right-hand side of (17) and solving for p, we require that

$$p > \frac{1+2\sigma\log n}{\log\frac{\rho}{\tau\gamma^2}}\,. \tag{18}$$

From (16) and (18) we thus see that it suffices to have

$$p > \frac{((1-\sigma)\alpha-\delta)\log n}{2C\log\frac{\gamma}{\tau}} \ge \frac{1+2\sigma\log n}{\log\frac{\rho}{\tau\gamma^2}}\,,$$

or equivalently,

$$\frac{\log\frac{\rho}{\tau\gamma^2}}{\log\frac{\gamma}{\tau}} \ge \frac{\frac{2C}{\log n}+4C\sigma}{(1-\sigma)\alpha-\delta}\,. \tag{19}$$

Let us derive a lower bound for the left-hand side of (19). Fix the constant $\gamma > 1$ so that $\log\gamma = -\frac{\epsilon\log\tau_{\max}}{100000}$. By our assumptions we have $\tau \le \tau_{\max}$ and $1-\log_\tau\rho \ge \epsilon$, so we have the lower bound

$$\frac{\log\frac{\rho}{\tau\gamma^2}}{\log\frac{\gamma}{\tau}} = \frac{\log\rho-\log\tau-2\log\gamma}{\log\gamma-\log\tau} = \frac{1-\log_\tau\rho+\frac{2\log\gamma}{\log\tau}}{1-\frac{\log\gamma}{\log\tau}} \ge \frac{\epsilon+\frac{2\log\gamma}{\log\tau_{\max}}}{1-\frac{\log\gamma}{\log\tau_{\max}}} > 0.99\epsilon\,.$$

Thus, (19) holds for all large enough n when we require

$$0.99\epsilon \ge \frac{4C\sigma}{(1-\sigma)\alpha-\delta}\,.$$

Since $\alpha\epsilon < 1$, we have that (19) holds when we set

$$\sigma = \frac{0.99\epsilon(\alpha-\delta)}{4C+1} \le \frac{0.99\epsilon(\alpha-\delta)}{4C+0.99\alpha\epsilon}\,.$$

We also observe that $(1-\sigma)\alpha-\delta > 0$, or equivalently, $\sigma < (\alpha-\delta)/\alpha$ holds for our choice of σ. This completes the parameterization of the algorithm.

Running time. Let us now analyze the running time of the algorithm. The first and second steps run in time $\tilde{O}(nD)$ since $p = O(\log n)$ by (16) and f originates from Theorem 3 and hence is explicit. From (15) and $n^{\sigma}-1 < s$, we have $nD \le 4n^{1+(1-\sigma)\alpha} \le 4n^{2-\sigma}$. Since (15) holds, the third step of the algorithm runs in time $O\big((n/s)^{2+\eta}\big)$ for any constant $\eta > 0$ that we are free to choose. Since $n/s \le 2n^{1-\sigma}$ for all large enough n, we can choose $\eta > 0$ so that $(2+\eta)(1-\sigma) \le 2-\sigma$. Thus, the first, second, and third steps together run in time $O(n^{2-\sigma})$. The fourth step runs in time $O(n^{2-\sigma}+qs^2d)$. Indeed, observe from (17) that the inequality (14) holds for at most q entries in Z. We have $qs^2d \le qn^{2\sigma+\delta}$, which completes the running time analysis and the proof of Theorem 4. ◀

4 Proof of Corollary 7

A useful variant of the Problem 1 asks for all outlier pairs of distinct vectors drawn from a *single* set $S \subseteq \{-1,1\}^d$ rather than two sets X, Y. We observe that the single-set variant reduces to $\lceil \log |S| \rceil$ instances of the two-set variant by numbering the vectors in S with binary numbers from 0 to $|S|-1$ and splitting S into two sets X_i, Y_i based on the value of the i^{th} bit for each $i = 0, 1, \ldots, \lceil \log |S| \rceil - 1$.

We will need the following bound due to Hoeffding which provides an exponentially small upper bound on the deviation of a sum of bounded independent random variables from its expectation.

▶ **Theorem 19** (Hoeffding [14, Theorem 2]). *Let $Z_1, Z_2, \ldots, Z_D$ be independent random variables satisfying $\ell_i \le Z_i \le u_i$ for all $1 \le i \le D$, and let $Z = \sum_{i=1}^{D} Z_i$. Then, for all $c > 0$, the following holds:*

$$\Pr\left(Z - \mathrm{E}[Z] \ge c\right) \le \exp\left(-\frac{2c^2}{\sum_{i=1}^{D}(u_i - \ell_i)^2}\right) . \tag{20}$$

▶ **Corollary 7.** *For any constants $0 < \delta < \alpha$, $C > 60$, $0 < \rho_{\max} < 1$, and $\kappa > 1$, there exists a deterministic algorithm that solves almost all instances of Problem 6 in time*

$$O\left(n^{2-\frac{0.99(1-1/\kappa)(\alpha-\delta)}{4C+1}}\right)$$

assuming the parameters n, d, ρ satisfy the two constraints

1. $5\rho^{-2\kappa} \log n \le d \le n^{\delta}$ *and*
2. $n^{-\Theta(1)} \le \rho \le \rho_{\max}$. [11]

Proof. We reduce to (the single-set version of) Problem 1 and apply Theorem 4. Towards this end, in Theorem 4 set $\epsilon = 1 - 1/\kappa$ and $\tau_{\max} = \rho_{\max}^{\kappa}$. Suppose we are given an instance of Problem 6 whose parameters n, d, ρ satisfy the constraints. Set $\tau = \rho^{\kappa}$. We observe that the constraints in Theorem 4 are satisfied since (i) $d \le n^{\delta}$ holds by assumption, (ii) $\tau \le \tau_{\max}$ holds since $\tau = \rho^{\kappa} \le \rho_{\max}^{\kappa}$, (iii) since $\kappa > 1$ is a constant and $\tau = \rho^{\kappa}$ we can satisfy the requirement that $\tau \ge n^{-\Theta(1)}$ for any desired constant hidden by the $\Theta(1)$ notation [12] by our assumption that $\rho \ge n^{-\Theta(1)}$, and (iv) $\log_{\tau} \rho = \frac{\log \rho}{\log \tau} = \frac{\log \rho}{\log \rho^{\kappa}} = 1/\kappa \le 1 - \epsilon$.

We claim that $q = 1$ for almost all instances of Problem 6 whose parameters satisfy the constraints in Corollary 7. Indeed, by the Hoeffding bound (20) and the union bound, the probability that some other pair than the planted pair in an instance has inner product that exceeds τd in absolute value is at most

$$2n^2 \exp\bigl(-\tau^2 d/2\bigr) \le 2n^2 \exp\bigl(-\rho^{2\kappa} \cdot 5\rho^{-2\kappa} \log n\bigr) = 2n^{-1/2} ,$$

so $q = 1$ with high probability as n increases. The claimed running time follows by substituting the chosen constants and $q = 1$ to (5). ◀

References

1 Thomas D. Ahle, Rasmus Pagh, Ilya Razenshteyn, and Francesco Silvestri. On the complexity of inner product similarity join. *arXiv*, abs/1510.02824, 2015.

[11] The constant hidden by the $\Theta(1)$ notation depends on the constants $\delta, \alpha, C, \rho_{\max}$ but not on the other parameters. For details consult the proof.

[12] Consult the proof of Theorem 4 for the details of this constant.

2 Josh Alman and Ryan Williams. Probabilistic polynomials and Hamming nearest neighbors. In *Proc. 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 136–150, Los Alamitos, CA, USA, 2015. IEEE Computer Society.

3 Noga Alon. Problems and results in extremal combinatorics – I. *Discrete Math.*, 273(1-3):31–53, 2003.

4 Alexandr Andoni, Piotr Indyk, Huy L. Nguyen, and Ilya Razenshteyn. Beyond locality-sensitive hashing. In *Proc. 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1018–1028, Philadelphia, PA, USA, 2014. Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611973402.76.

5 Alexandr Andoni and Ilya Razenshteyn. Optimal data-dependent hashing for approximate near neighbors. In *Proc. 47th ACM Annual Symposium on the Theory of Computing (STOC)*, pages 793–801, New York, NY, USA, 2015. Association for Computing Machinery. doi:10.1145/2746539.2746553.

6 L. Elisa Celis, Omer Reingold, Gil Segev, and Udi Wieder. Balls and bins: Smaller hash families and faster evaluation. *SIAM J. Comput.*, 42(3):1030–1050, 2013.

7 Timothy M. Chan and Ryan Williams. Deterministic APSP, orthogonal vectors, and more: Quickly derandomizing Razborov-Smolensky. In Robert Krauthgamer, editor, *Proc. 27th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1246–1255, Arlington, VA, USA, 2016. Society for Industrial and Applied Mathematics.

8 Moshe Dubiner. Bucketing coding and information theory for the statistical high-dimensional nearest-neighbor problem. *IEEE Trans. Inf. Theory*, 56(8):4166–4179, 2010. doi:10.1109/TIT.2010.2050814.

9 Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM J. Comput.*, 39(2):606–645, 2009. doi:10.1137/070684914.

10 Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In Malcolm P. Atkinson, Maria E. Orlowska, Patrick Valduriez, Stanley B. Zdonik, and Michael L. Brodie, editors, *Proc. 25th International Conference on Very Large Data Bases (VLDB'99)*, pages 518–529, Edinburgh, Scotland, UK, 1999. Morgan Kaufmann.

11 Parikshit Gopalan, Daniek Kane, and Raghu Meka. Pseudorandomness via the Discrete Fourier Transform. In *Proc. IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 903–922, Berkeley, CA, USA, 2015. IEEE Computer Society.

12 Parikshit Gopalan, Raghu Meka, Omer Reingold, and David Zuckerman. Pseudorandom generators for combinatorial shapes. *SIAM J. Comput.*, 42(3):1051–1076, 2013.

13 Elena Grigorescu, Lev Reyzin, and Santosh Vempala. On noise-tolerant learning of sparse parities and related problems. In *Proc. 22nd International Conference on Algorithmic Learning Theory (ALT)*, pages 413–424, Berlin, Germany, 2011. Springer. doi:10.1007/978-3-642-24412-4_32.

14 Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

15 Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bull. Amer. Math. Soc.*, 43(4):439–561, 2006.

16 Russell Impagliazzo and Ramamohan Paturi. On the complexity of k-SAT. *J. Comput. Syst. Sci.*, 62(2):367–375, 2001. doi:10.1006/jcss.2000.1727.

17 Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. 30th Annual ACM Symposium on the Theory of Computing (STOC)*, pages 604–613, New York, NY, USA, 1998. Association for Computing Machinery. doi:10.1145/276698.276876.

18 Daniel M. Kane, Raghu Meka, and Jelani Nelson. Almost optimal explicit Johnson-Lindenstrauss families. In *Proc. 14th International Workshop on Approximation, Randomization, and Combinatorial Optimization, RANDOM and 15th International Workshop on Algorithms and Techniques, APPROX*, pages 628–639, Princeton, NJ, USA, 2011.

19 Michael Kapralov. Smooth tradeoffs between insert and query complexity in nearest neighbor search. In *Proc. 34th ACM Symposium on Principles of Database Systems (PODS)*, pages 329–342, New York, NY, USA, 2015. Association for Computing Machinery. doi:10.1145/2745754.2745761.

20 Matti Karppa, Petteri Kaski, and Jukka Kohonen. A faster subquadratic algorithm for finding outlier correlations. In *Proc. 27th Annual ACM-SIAM Symposium on Discrete Algorithms, (SODA)*, pages 1288–1305, Arlington, VA, USA, 2016. Society for Industrial and Applied Mathematics.

21 Pravesh K. Kothari and Raghu Meka. Almost optimal pseudorandom generators for spherical caps. In *Proc. 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 247–256, Portland, OR, USA, 2015.

22 François Le Gall. Faster algorithms for rectangular matrix multiplication. In *Proc. 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 514–523, Los Alamitos, CA, USA, 2012. IEEE Computer Society. doi:10.1109/FOCS.2012.80.

23 A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan graphs. *Combinatorica*, 8:261–277, 1988.

24 Alexander May and Ilya Ozerov. On computing nearest neighbors with applications to decoding of binary linear codes. In *Proc. EUROCRYPT 2015 - 34th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 203–228, Berlin, Germany, 2015. Springer. doi:10.1007/978-3-662-46800-5_9.

25 Elchanan Mossel, Ryan O'Donnell, and Rocco A. Servedio. Learning functions of k relevant variables. *J. Comput. Syst. Sci.*, 69(3):421–434, 2004. doi:10.1016/j.jcss.2004.04.002.

26 Rajeev Motwani, Assaf Naor, and Rina Panigrahy. Lower bounds on locality sensitive hashing. *SIAM J. Discrete Math.*, 21(4):930–935, 2007. doi:10.1137/050646858.

27 Ryan O'Donnell, Yi Wu, and Yuan Zhou. Optimal lower bounds for locality-sensitive hashing (except when q is tiny). *ACM Trans. Comput. Theory*, 6(1):Article 5, 2014. doi:10.1145/2578221.

28 Rasmus Pagh. Locality-sensitive hashing without false negatives. In *Proc. 27th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1–9, Philadelphia, PA, USA, 2016. Society for Industrial and Applied Mathematics.

29 Ramamohan Paturi, Sanguthevar Rajasekaran, and John H. Reif. The light bulb problem. In *Proc. 2nd Annual Workshop on Computational Learning Theory (COLT)*, pages 261–268, New York, NY, USA, 1989. Association for Computing Machinery.

30 Ninh Pham and Rasmus Pagh. Scalability and total recall with fast CoveringLSH. *arXiv*, abs/1602.02620, 2016.

31 Omer Reingold, Salil Vadhan, and Avi Wigderson. Entropy waves, the zig-zag graph product, and new constant-degree expanders. *Ann. of Math.*, 155(1):157–187, 2002.

32 Victor Shoup. New algorithms for finding irreducible polynomials over finite fields. *Math. Comp.*, 54:435–447, 1990.

33 Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem. *J. ACM*, 62(2):Article 13, 2015. doi:10.1145/2728167.

34 Leslie G. Valiant. Functionality in neural nets. In *Proc. 1st Annual Workshop on Computational Learning Theory (COLT)*, pages 28–39, New York, NY, USA, 1988. Association for Computing Machinery.

A An expander family

This section proves Lemma 9 following Reingold, Vadhan and Wigderson [31]; we present the proof for completeness of exposition only with no claim of originality. Following Reingold, Vadhan and Wigderson [31] we will work with *normalized* eigenvalues. To avoid confusion with the unnormalized treatment in the manuscript proper, we say that a graph is a $[D,\Delta,\lambda]$-*graph* if the graph has D vertices, is Δ-regular, and $|\lambda_2|/\Delta \le \lambda$. (Here $|\lambda_2|$ is the unnormalized second eigenvalue as defined in the manuscript proper.)

We refer to Sections 2.3 and 3.1 of Reingold, Vadhan, and Wigderson [31] for the definition of the square G^2 of a graph G, the tensor product $G_1 \otimes G_2$ of graphs G_1, G_2, and the zigzag product $G \circledz H$ of graphs G, H. The following omnibus result collects elements of Propositions 2.3, Proposition 2.4, Theorem 3.2 and Theorem 4.3 of [31] which will be sufficient to control the second normalized eigenvalue for our present purposes. (We choose to omit the details of the rotation maps with the understanding that they can be found in [31].)

▶ **Lemma 20** (Reingold, Vadhan, and Wigderson [31]). *The following bounds hold.*

1. *If G is a $[D,\Delta,\lambda]$-graph, then G^2 is a $[D,\Delta^2,\lambda^2]$-graph.*
2. *If G_1 is a $[D_1,\Delta_1,\lambda_1]$-graph and G_2 is a $[D_2,\Delta_2,\lambda_2]$-graph, then $G_1 \otimes G_2$ is a $[D_1D_2,\Delta_1\Delta_2,\max(\lambda_1,\lambda_2)]$-graph.*
3. *If G is a $[D_1,\Delta_1,\lambda_1]$-graph and H a $[\Delta_1,\Delta_2,\lambda_2]$-graph, then $G \circledz H$ is a $[D_1\Delta_1,\Delta_2^2,f(\lambda_1,\lambda_2)]$-graph with*

$$f(\lambda_1,\lambda_2) = \frac{1}{2}\left(1-\lambda_2^2\right)\lambda_1 + \frac{1}{2}\sqrt{\left(1-\lambda_2^2\right)^2\lambda_1^2 + 4\lambda_2^2} \le \lambda_1 + \lambda_2 \,.$$

Let us study the following sequence of graphs. Let H be a $[D,\Delta,\lambda]$-graph. Let $G_1 = H^2$, $G_2 = H \otimes H$, and for $t = 3,4,\ldots$ let

$$G_t = \left(G_{\lceil\frac{t-1}{2}\rceil} \otimes G_{\lfloor\frac{t-1}{2}\rfloor}\right)^2 \circledz H \,. \tag{21}$$

From Lemma 20 it is easily seen that G_t is a $[D^t,\Delta^2,\lambda_t]$-graph with λ_t defined by

$$\begin{aligned} \lambda_1 &= \lambda^2\,, & \\ \lambda_2 &= \lambda\,, & \\ \lambda_{2t-1} &= \lambda + \lambda_{t-1}^2\,, & &\text{for } t = 2,3\ldots\,, \text{ and} \\ \lambda_{2t} &= \max(\lambda+\lambda_t^2, \lambda+\lambda_{t-1}^2)\,, & &\text{for } t = 2,3,\ldots\,. \end{aligned}$$

▶ **Lemma 21** (Reingold, Vadhan, and Wigderson [31, Theorem 3.3]). *The rotation map* Rot_{G_t} *can be computed in time* $\mathrm{poly}(t,\log D)$ *and by making* $\mathrm{poly}(t)$ *evaluations of* Rot_H*.*

▶ **Lemma 22.** *If $0 \le \lambda \le 1/4$ then $\lambda_t \le \lambda + 4\lambda^2$ for all $t \ge 1$.*

Proof. The conclusion is immediate for $t \le 2$. So suppose that the conclusion holds up to $2t-2$. We need to show that the conclusion holds for λ_{2t-1} and λ_{2t}. By induction, it suffices to show that

$$\lambda_{2t-1} \le \lambda + (\lambda+4\lambda^2)^2 \le \lambda + 4\lambda^2 \,.$$

Observing that $\lambda^2 + 8\lambda^3 + 16\lambda^4 \le 4\lambda^2$ holds for $0 \le \lambda \le 1/4$ yields the desired conclusion. The proof for λ_{2t} is identical. ◀

Finally, we construct the expanders that we require in the manuscript proper.

▶ **Lemma 23** (Lemma 9 stated with normalized eigenvalue notation). *For all integers $t \geq 1$ and $b \geq 10$ there exists a $[2^{16bt}, 2^{4b}, 16 \cdot 2^{-b}]$-graph whose rotation map can be evaluated in time* $\mathrm{poly}(b, t)$.

Proof. Take $q = 2^b$ and $d = 15$ in Proposition 5.3 of Reingold, Vadhan, and Wigderson [31] to obtain a $[2^{16b}, 2^{2b}, 15 \cdot 2^{-b}]$-graph H whose rotation map can be computed in time $\mathrm{poly}(b)$. (Indeed, observe that an irreducible polynomial to perform the required arithmetic in the finite field of order 2^b can be constructed in deterministic time $\mathrm{poly}(b)$ by an algorithm of Shoup [32].) Let us study the sequence G_t given by (21). The time complexity of the rotation map follows immediately from Lemma 21. Since $b \geq 10$, Lemma 22 gives that $\lambda_t \leq \lambda + 4\lambda^2$ for all $t \geq 1$. Take $\lambda = 15 \cdot 2^{-b}$ and observe that since $b \geq 10$ we have $2^{-b} < 1/900$. Thus, $\lambda_t \leq 15 \cdot 2^{-b} + 4(15 \cdot 2^{-b})^2 = 15 \cdot 2^{-b} + 900 \cdot 2^{-2b} \leq 16 \cdot 2^{-b}$. ◀