Binary embeddings with structured hashed projections (Supplementary Material)

In this section we prove Theorem 4.1 and Theorem 4.2. We will use notation from Lemma 4.1.

7. Proof of Theorem 4.1

We start with the following technical lemma:

Lemma 7.1. *Let $\{Z_1, ..., Z_k\}$ be the set of k independent random variables defined on Ω such that each Z_i has the same distribution and $Z_i \in \{0, 1\}$. Let $\{\mathcal{F}_1, ..., \mathcal{F}_k\}$ be the set of events, where each $\mathcal{F}_i$ is in the σ-field defined by Z_i (in particular $\mathcal{F}_i$ does not depend on the σ-field $\sigma(Z_1, ..., Z_{i-1}, Z_{i+1}, ...Z_k)$). Assume that there exists $\mu < \frac{1}{2}$ such that: $\mathbb{P}(\mathcal{F}_i) \leq \mu$ for $i = 1, ..., k$. Let $\{U_1, ..., U_k\}$ be the set of k random variables such that $U_i \in \{0, 1\}$ and $U_i|\mathcal{F}_i = Z_i|\mathcal{F}_i$ for $i = 1, ..., k$, where $X|\mathcal{F}$ stands for the random variable X truncated to the event $\mathcal{F}$. Assume furthermore that $E(U_i) = E(Z_i)$ for $i = 1, ..., k$. Denote $U = \frac{U_1+...+U_k}{k}$. Then the following is true.*

$$\mathbb{P}(|U-EU| > \epsilon) \leq \frac{1}{\pi} \sum_{r=\frac{\epsilon k}{2}}^{k} \frac{1}{\sqrt{r}} (\frac{ke}{r})^r \mu^r (1-\mu)^{k-r} + 2e^{-\frac{\epsilon^2 k}{2}}. \tag{6}$$

Proof. Let us consider the event $\mathcal{F}_{bad} = \mathcal{F}_1 \cup ... \cup \mathcal{F}_k$. Note that $\mathcal{F}_{bad}$ may be represented by the union of the so-called r-blocks, i.e.

$$\mathcal{F}_{bad} = \bigcup_{Q \subseteq \{1,...,k\}} (\bigcap_{q \in Q} \mathcal{F}_q \bigcap_{q \in \{1,...,k\} \setminus Q} \mathcal{F}_q^c), \tag{7}$$

where $\mathcal{F}^c$ stands for the complement of event $\mathcal{F}$. Let us fix now some $Q \subseteq \{1, ..., k\}$. Denote

$$\mathcal{F}_Q = \bigcap_{q \in Q} \mathcal{F}_q \bigcap_{q \in \{1,...,k\} \setminus Q} \mathcal{F}_q^c. \tag{8}$$

note that $\mathbb{P}(\mathcal{F}_Q) \leq \mu^r (1-\mu)^{k-r}$. It follows directly from the Bernoulli scheme.

Denote $Z = \frac{Z_1+...+Z_k}{k}$. From what we have just said and from the definition of $\{\mathcal{F}_1, ..., \mathcal{F}_k\}$ we conclude that for any given c the following holds:

$$\mathbb{P}(|U - Z| > c) \leq \sum_{r=ck}^{k} \binom{k}{r} \mu^r (1-\mu)^{k-r}. \tag{9}$$

Note also that from the assumptions of the lemma we trivially get: $E(U) = E(Z)$.

Let us consider now the expression $\mathbb{P}(|U - E(U)|) > \epsilon$.

We get: $\mathbb{P}(|U - E(U)| > \epsilon) = \mathbb{P}(|U - E(Z)| > \epsilon) = \mathbb{P}(|U - Z + Z - E(Z)| > \epsilon) \leq \mathbb{P}(|U - Z| + |Z - E(Z)| > \epsilon) \leq \mathbb{P}(|U - Z| > \frac{\epsilon}{2}) + \mathbb{P}(|Z - E(Z)| > \frac{\epsilon}{2})$.

From 9 we get:

$$\mathbb{P}(|U - Z| > \frac{\epsilon}{2}) \leq \sum_{r=\frac{\epsilon}{2}}^{k} \binom{k}{r} \mu^r (1-\mu)^{k-r}. \tag{10}$$

Let us consider now the expression:

$$\xi = \sum_{r=\frac{\epsilon k}{2}}^{k} \binom{k}{r} \mu^r (1-\mu)^{k-r}. \tag{11}$$

We have:

$$\begin{aligned} \xi &\leq \sum_{r=\frac{\epsilon k}{2}}^{k} \frac{(k-r+1)...(k)}{r!} \mu^r (1-\mu)^{k-r} \\ &\leq \sum_{r=\frac{\epsilon k}{2}}^{k} \frac{k^r}{r!} \mu^r (1-\mu)^{k-r} \end{aligned} \tag{12}$$

From the Stirling's formula we get: $r! = \frac{2\pi r^{r+\frac{1}{2}}}{e^r}(1 + o_r(1))$. Thus we obtain:

$$\begin{aligned} \xi &\leq (1 + o_r(1)) \sum_{r=\frac{\epsilon k}{2}}^{k} \frac{k^r e^r}{2\pi r^{r+\frac{1}{2}}} \mu^r (1-\mu)^{k-r} \\ &\leq \frac{1}{\pi} \sum_{r=\frac{\epsilon k}{2}}^{k} \frac{1}{\sqrt{r}} (\frac{ke}{r})^r \mu^r (1-\mu)^{k-r} \end{aligned} \tag{13}$$

for r large enough.

Now we will use the following version of standard Azuma's inequality:

Lemma 7.2. *Let $W_1, ..., W_k$ be k independent random variables such that $E(W_1) = ...E(W_k) = 0$. Assume that $-\alpha_i \leq W_{i+1} - W_i \leq \beta_i$ for $i = 2, ..., k-1$. Then the following is true:*

$$\mathbb{P}(|\sum_{i=1}^{k} W_i| > a) \leq 2e^{-\frac{2a^2}{\sum_{i=1}^{k} (\alpha_i + \beta_i)^2}}$$

Now, using Lemma 7.2 for $W_i = X_i - E(X_i)$ and $\alpha_i = E(X_i), \beta_i = 1 - E(X_i)$ we obtain:

$$\mathbb{P}(|X - EX| > \frac{a}{2}) \leq 2e^{-\frac{a^2 k}{2}}. \tag{14}$$

Combining 13 and 14, we obtain the statement of the lemma.

□

Our next lemma explains the role the Hadamard matrix plays in the entire extended Ψ-regular hashing mechanism.

Lemma 7.3. *Let n denote data dimensionality and let $f(n)$ be an arbitrary positive function. Let D be the set of all L_2-normalized data points, where no two data points are identical. Assume that $|D| = N$. Consider the $\binom{N}{2}$ hyperplanes $H_{p,r}$ spanned by pairs of different vectors $\{p, r\}$ from D. Then after applying linear transformation $\mathcal{HR}$ each hyperplane $H_{p,r}$ is transformed into another hyperplane $H_{p,r}^{\mathcal{HR}}$. Furthermore, the probability $\mathcal{P}_{\mathcal{HR}}$ that for every $H_{p,r}^{\mathcal{HR}}$ there exist two orthonormal vectors $x = (x_1, ..., x_n), y = (y_1, ..., y_n)$ in $H_{p,r}^{\mathcal{HR}}$ such that: $|x_i|, |y_i| \leq \frac{f(n)}{\sqrt{n}}$ satisfies:*

$$\mathcal{P}_{\mathcal{HR}} \geq 1 - 4\binom{N}{2} e^{-\frac{f^2(n)}{2}}.$$

Proof. We have already noted in the proof of Lemma 4.1 that $\mathcal{HR}$ is an orthogonal matrix. Thus, as an isometry, it clearly transforms each 2-dimensional hyperplane into another 2-dimensional hyperplane. For every pair $\{p, r\}$, let us consider an arbitrary fixed orthonormal pair $\{u, v\}$ spanning $H_{p,r}$. Denote $u = (u_1, ..., u_n)$. Let us denote by $u^{\mathcal{HR}}$ vector obtained from u after applying transformation $\mathcal{HR}$. Note that the j^{th} coordinate of $u^{\mathcal{HR}}$ is of the form:

$$u_j^{\mathcal{HR}} = u_1 T_1 + ... + u_n T_n, \tag{15}$$

where $T_1, ..., T_n$ are independent random variables satisfying:

$$T_i = \begin{cases} \frac{1}{\sqrt{n}} & \text{w.p } \frac{1}{2}, \\ -\frac{1}{\sqrt{n}} & \text{otherwise.} \end{cases} \tag{16}$$

The latter comes straightforwardly from the form of the L_2-normalized Hadamard matrix (i.e a Hadamard matrix, where each row and column is L_2-normalized).

But then, from Lemma 7.2, and the fact that $\|u\|_2 = 1$, we get for any $a > 0$:

$$\mathbb{P}(|u_1 T_1 + ... + u_n T_n| \geq a) \leq 2e^{-\frac{2a^2}{\sum_{i=1}^n (2u_i)^2}} \leq 2e^{-\frac{a^2}{2}}. \tag{17}$$

Similar analysis is correct for $v^{\mathcal{HR}}$. Note that $v^{\mathcal{HR}}$ is orthogonal to $u^{\mathcal{HR}}$ since v and u are orthogonal. Furthermore, both $v^{\mathcal{HR}}$ and $u^{\mathcal{HR}}$ are L_2-normalized. Thus $\{u^{\mathcal{HR}}, v^{\mathcal{HR}}\}$ is an orthonormal pair.

To complete the proof, it suffices to take $a = f(n)$ and apply the union bound over all vectors $u^{\mathcal{HR}}$, $v^{\mathcal{HR}}$ for all $\binom{N}{2}$ hyperplanes. □

From the lemma above we see that applying Hadamard matrix enables us to assume with high probability that for every hyperplane $H_{p,r}$ there exists an orthonormal basis consisting of vectors with elements of absolute values at most $\frac{f(n)}{\sqrt{n}}$. We call this event $\mathcal{E}_f$. Note that whether $\mathcal{E}_f$ holds or not is determined only by $\mathcal{H}$, $\mathcal{R}$ and the initial dataset D.

Let us proceed with the proof of Theorem 4.1. Let us assume that event $\mathcal{E}_f$ holds. Without loss of generality we may assume that we have the short Ψ-regular hashing mechanism with an extra property that every $H_{p,r}$ has an orthonormal basis consisting of vectors with elements of absolute value at most $\frac{f(n)}{\sqrt{n}}$. Fix two vectors p, r from the dataset D. Denote by $\{x, y\}$ the orthonormal basis of $H_{p,r}$ with the above property. Let us fix the ith row of $\mathcal{P}$ and denote it as $(p_{i,1}, ..., p_{i,n})$. After being multiplied by the diagonal matrix $\mathcal{D}$ we obtain another vector:

$$w = (\mathcal{P}_{i,1} d_1, ..., \mathcal{P}_{i,n} d_n), \tag{18}$$

where:

$$\mathcal{D}_{i,j} = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix}. \tag{19}$$

We have already noted that in the proof of Lemma 4.1 that it is the projection of w into $H_{p,r}$ that determines whether the value of the associated random variable X_i is 0 or 1. To be more specific, we showed that $X_i = 1$ iff the projection is in the region $\mathcal{U}_{p,r}$. Let us write down the coordinates of the projection of w into $H_{p,r}$ in the $\{x, y\}$-coordinate system. The coordinates are the dot-products of w with x and y respectively thus in the $\{x, y\}$-coordinate system we can write w as:

$$w_{\{x,y\}} = (\mathcal{P}_{i,1} d_1 x_1, ..., \mathcal{P}_{i,n} d_n x_n, \mathcal{P}_{i,1} d_1 y_1, ..., \mathcal{P}_{i,n} d_n y_n). \tag{20}$$

Note that both coordinates are Gaussian random variables and they are independent since they were constructed by projecting a Gaussian vector into two orthogonal vectors. Now note that from our assumption about the structure of $\mathcal{P}$ we can conclude that both coordinates may be represented as sums of weighted Gaussian random variables g_i for $i = 1, ..., t$, i.e.:

$$w_{\{x,y\}} = (g_1 s_{i,1} + ... + g_t s_{i,t}, g_1 v_{i,1} + ... + g_t v_{i,t}), \tag{21}$$

where each $s_{i,j}, v_{i,j}$ is of the form $d_z x_z$ or $d_z y_z$ for some z that depends only on i, j. Note also that

$$s_{i,1}^2 + ... + s_{i,t}^2 = v_{i,1}^2 + ... + v_{i,t}^2. \tag{22}$$

The latter inequality comes from the fact that, by 20, both coordinates of $w_{\{x,y\}}$ have the same distribution.

Let us denote $s_i = (s_{i,1}, ..., s_{i,t})$, $v_i = (v_{i,1}, ..., v_{i,t})$ for $i = 1, ..., k$. We need the following lemma stating that with high probability vectors $s_1, ..., s_k, v_1, ..., v_k$ are close to be pairwise orthogonal.

Lemma 7.4. *Let us assume that $\mathcal{E}_f$ holds. Let $f(n)$ be an arbitrary positive function. Then for every $a > 0$ with probability at least $\mathbb{P}_{succ} \geq 1 - 4\binom{k}{2} e^{-\frac{2a^2 n}{f^4(n)}}$, taken under coin tosses used to construct $\mathcal{D}$, the following is true for every $1 \leq i_1 \neq i_2 \leq k$:*

$$|\sum_{u=1}^{n} s_{i_1,u} v_{i_1,u}| \leq a\chi(\mathcal{P}) + \Psi \frac{f^2(n)}{n},$$

$$|\sum_{u=1}^{n} s_{i_1,u} s_{i_2,u}| \leq a\chi(\mathcal{P}) + \Psi \frac{f^2(n)}{n},$$

$$|\sum_{u=1}^{n} v_{i_1,u} v_{i_2,u}| \leq a\chi(\mathcal{P}) + \Psi \frac{f^2(n)}{n},$$

$$|\sum_{u=1}^{n} s_{i_1,u} v_{i_2,u}| \leq a\chi(\mathcal{P}) + \Psi \frac{f^2(n)}{n}.$$

Proof. Note that the we get the first inequality for free from the fact that x is orthogonal to y (in other words, $\sum_{u=1}^{n} s_{i_1,u} v_{i_1,u}$ can be represented as $C \sum_{u=1}^{n} x_i y_i$ and the latter expression is clearly 0). Let us consider now one of the three remaining expressions. Note that they can be rewritten as:

$$E = \sum_{i=1}^{n} d_{\rho(i)} d_{\lambda(i)} x_{\zeta(i)} x_{\gamma(i)} \tag{23}$$

or

$$E = \sum_{i=1}^{n} d_{\rho(i)} d_{\lambda(i)} y_{\zeta(i)} y_{\gamma(i)} \tag{24}$$

or

$$E = \sum_{i=1}^{n} d_{\rho(i)} d_{\lambda(i)} x_{\zeta(i)} y_{\gamma(i)} \tag{25}$$

for some $\rho, \lambda, \zeta, \gamma$. Note also that from the Ψ-regularity condition we immediately obtain that $\rho(i) = \lambda(i)$ for at most Ψ elements of each sum. Get rid of these elements from each sum and consider the remaining ones. From the definition of the $\mathcal{P}$-chromatic number, those remaining ones can be partitioned into at most $\chi(\mathcal{P})$ parts, each consisting of elements that are independent random variables (since in the corresponding graph there are no edges between them). Thus, for the sum corresponding to each part one can apply Lemma 7.2. Thus one can conclude that the sum differs from its expectation (which clearly is 0 since $E(d_i d_j) = 0$ for $i \neq j$) by a with probability at most:

$$\mathbb{P}_a \leq 2e^{-\frac{2a^2}{\sum_{i=1}^{n} x_{\zeta(i)} x_{\gamma(i)}}}, \tag{26}$$

or

$$\mathbb{P}_a \leq 2e^{-\frac{2a^2}{\sum_{i=1}^{n} y_{\zeta(i)} y_{\gamma(i)}}}, \tag{27}$$

or

$$\mathbb{P}_a \leq 2e^{-\frac{2a^2}{\sum_{i=1}^{n} x_{\zeta(i)} y_{\gamma(i)}}}. \tag{28}$$

Now it is time to use the fact that event $\mathcal{E}_f$ holds. Then we know that: $|x_i|, |y_i| \leq \frac{f(n)}{\sqrt{n}}$ for $i = 1, ..., n$. Substituting this upper bound for $|x_i|, |y_i|$ in the derived expressions on the probabilities coming from Lemma 7.2, and then taking the union bound, we complete the proof. □

We can finish the proof of Theorem 4.1. From Lemma 7.4 we see that $s_1, ..., s_k, v_1, ..., v_k$ are close to pairwise orthogonal with high probability. Let us fix some positive function $f(n) > 0$ and some $a > 0$. Denote

$$\Delta = a\chi(\mathcal{P}) + \Psi \frac{f^2(n)}{n}. \tag{29}$$

Note that, by Lemma 7.4 we see that applying Gram-Schmidt process we can obtain a system of pairwise orthogonal vectors $\tilde{s}_1, ..., \tilde{s}_k, \tilde{v}_1, ..., \tilde{v}_k$ such that

$$\|\tilde{v}_i - v_i\|_2 \leq \sigma(k)\Delta. \tag{30}$$

and

$$\|\tilde{s}_i - s_i\|_2 \leq \sigma(k)\Delta, \tag{31}$$

where $\sigma(k) > 0$ is some function of k (it does not depend on n and t). Note that for n, t large enough we have: $\sigma(k)\Delta \leq \sqrt{a}\chi(\mathcal{P}) + \Psi \frac{f^2(n)}{\sqrt{n}}$.

Let us consider again $w_{x,y}$. Replacing s_i by $\tilde{s}_i$ and v_i by $\tilde{v}_i$ in the formula on $w_{x,y}$, we obtain another Gaussian vector: $\tilde{w}_{x,y}$ for each row i of the matrix $\mathcal{P}$. Note however that vectors $\tilde{w}_{x,y}$ have one crucial advantage over vectors $w_{x,y}$, namely they are independent. That comes from the fact that $\tilde{s}_1, ..., \tilde{s}_k, \tilde{v}_1, ..., \tilde{v}_k$ are pairwise orthogonal. Note also that from 36 and 37 we obtain that the angular distance between $w_{x,y}$ and $\tilde{w}_{x,y}$ is at most $\sigma(k)\Delta$.

Let Z_i for $i = 1, ...k$ be an indicator random variable that is zero if $\tilde{w}_{x,y}$ is inside the region $\mathcal{U}_{p,r}$ and zero otherwise. Let U_i for $i = 1, ...k$ be an indicator random variable that is zero if $w_{x,y}$ is inside the region $\mathcal{U}_{p,r}$ and zero otherwise. Note that $\tilde{\theta}_{p,r}^{n} = \frac{U_1 + ... + U_k}{k}$. Furthermore, random

variables $Z_1, ..., Z_k, U_1, ..., U_k$ satisfy the assumptions of Lemma 7.1 with $\mu \leq \frac{8\tau}{\theta_{p,r}}$, where $\tau = \sigma(k)\Delta$. Indeed, random variables Z_i are independent since vectors $\tilde{w}_{x,y}$ are independent. From what we have said so far we know that each of them takes value one with probability exactly $\frac{\theta_{p,r}}{\pi}$. Furthermore $Z_i \neq U_i$ only if $w_{x,y}$ is inside $\mathcal{U}_{p,r}$ and $\tilde{w}_{x,y}$ is outside $\mathcal{U}_{p,r}$ or vice versa. The latter event implies (thus it is included in the event) that $w_{x,y}$ is near the border of the region $\mathcal{U}_{p,r}$, namely within an angular distance $\frac{\epsilon}{\theta_{p,r}}$ from one of the four semi-lines defining $\mathcal{U}_{p,r}$. Thus in particular, an event $Z_i \neq U_i$ is contained in the event of probability at most $2 \cdot 4 \cdot \frac{\epsilon}{\theta_{p,r}}$ that depends only on one $w_{x,y}$.

But then we can apply Lemma 7.1. All we need is to assume that the premises of Lemma 7.4 are satisfied. But this is the case with probability specified in Lemma 7.3 and this probability is taken under random coin tosses used to product $\mathcal{H}$ and $\mathcal{R}$, thus independently from the random coin tosses used to produce $\mathcal{D}$. Putting it all together we obtain the statement of Theorem 4.1.

8. Proof of Theorem 4.2

We will borrow some notation from the proof of Theorem 4.1. Note however that in this setting no preprocessing with the use of matrices $\mathcal{H}$ and $\mathcal{R}$ is applied.

Lemma 8.1. *Define $U_1, ..., U_k$ as in the proof of Theorem 4.1. Assume that the following is true:*

$$|\sum_{u=1}^{n} s_{i_1,u} v_{i_1,u}| \leq \Delta,$$

$$|\sum_{u=1}^{n} s_{i_1,u} s_{i_2,u}| \leq \Delta,$$

$$|\sum_{u=1}^{n} v_{i_1,u} v_{i_2,u}| \leq \Delta,$$

$$|\sum_{u=1}^{n} s_{i_1,u} v_{i_2,u}| \leq \Delta.$$

for some $0 < \Delta < 1$. The the following is true for every fixed $1 \leq i < j \leq k$:

$$|\mathbb{P}(U_i U_j = 1) - \mathbb{P}(U_i = 1)\mathbb{P}(U_j = 1)| = O(\Delta).$$

The lemma follows from the exactly the same analysis that was done in the last section of the proof of Theorem 4.1 thus we leave it to the reader as an exercise.

Note that we have:

$$Var(\tilde{\theta}^n_{p,r}) = Var(\frac{U_1 + ... + U_k}{k}) = \frac{1}{k^2}(\sum_{i=1}^{k} Var(U_i) + \sum_{i \neq j} Cov(U_i, U_j)). \quad (32)$$

Since U_i is an indicator random variable that takes value one with probability $\frac{\theta_{p,r}}{\pi}$, we get:

$$Var(U_i) = E(U_i^2) - E(U_i)^2 = \frac{\theta_{p,r}}{\pi}(1 - \frac{\theta_{p,r}}{\pi}). \quad (33)$$

Thus we have:

$$Var(\tilde{\theta}^n_{p,r}) = \frac{1}{k}\frac{\theta_{p,r}(\pi - \theta_{p,r})}{\pi^2} + \frac{1}{k^2}\sum_{i \neq j} Cov(U_i, U_j). \quad (34)$$

Note however that $Cov(U_i, U_j)$ is exactly: $\mathbb{P}(U_i U_j = 1) - \mathbb{P}(U_i = 1)\mathbb{P}(U_j = 1)$.

Therefore, using Lemma 8.1, we obtain:

$$Var(\tilde{\theta}^n_{p,r}) = \frac{1}{k}\frac{\theta_{p,r}(\pi - \theta_{p,r})}{\pi^2} + O(\Delta). \quad (35)$$

It suffices to estimate parameter Δ. We proceed as in the previous proof. We only need to be a little bit more cautious since the condition: $|x_i|, |y_i| \leq \frac{f(n)}{\sqrt{n}}$ cannot be assumed right now. We select two rows: i_1, i_2 of $\mathcal{P}$. Note that again we see that applying Gram-Schmidt process, we can obtain a system of pairwise orthogonal vectors $\tilde{s}_{i_1}, \tilde{s}_{i_i}, \tilde{v}_{i_i}, \tilde{v}_{i_2}$ such that

$$\|\tilde{v}_{i_1} - v_{i_2}\|_2 = O(\Delta), \quad (36)$$

and

$$\|\tilde{s}_{i_1} - s_{i_2}\|_2 = O(\Delta). \quad (37)$$

The fact that right now the above upper bounds are not multiplied by k, as it was the case in the previous proof, plays key role in obtaining nontrivial concentration results even when no Hadamard mechanism is applied.

We consider the related sums: $E_1 = \sum_{i=1}^{n} d_{\rho(i)} d_{\lambda(i)} x_{\zeta(i)} x_{\gamma(i)}, E_2 = \sum_{i=1}^{n} d_{\rho(i)} d_{\lambda(i)} y_{\zeta(i)} y_{\gamma(i)}, E_3 = \sum_{i=1}^{n} d_{\rho(i)} d_{\lambda(i)} x_{\zeta(i)} y_{\gamma(i)}$ as before. We can again partition each sum into at most $\chi(\mathcal{P})$ sub-chunks, where this time $\chi(\mathcal{P}) \leq 3$ (since $\mathcal{P}$ is Toeplitz Gaussian). The problem is that applying Lemma 7.2, we get bounds that depend on the expressions of the form

$$\alpha_{x,i} = \sum_{j=1}^{n} x_j^2 x_{j+i}^2, \quad (38)$$

Table 2: Mean and std of the test error versus the size of the hash (k) / size of the reduction (n/k) for the network.

k / $\frac{n}{k}$	Circulant [%]	Random [%]	BinPerm [%]	BinCirc [%]	HalfShift [%]	Toeplitz [%]	VerHorShift [%]
1024 / 1	3.53 ± 0.16	2.78 ± 0.10	3.69 ± 0.21	6.79 ± 0.49	3.54 ± 0.16	3.16 ± 0.19	3.74 ± 0.16
512 / 2	5.42 ± 0.83	3.61 ± 0.19	4.68 ± 0.35	8.10 ± 1.85	5.13 ± 2.15	4.97 ± 0.53	5.55 ± 0.62
256 / 4	11.56 ± 1.42	4.79 ± 0.13	7.43 ± 1.31	6.13 ± 1.42	5.98 ± 1.05	9.48 ± 1.88	10.96 ± 2.78
128 / 8	22.10 ± 5.42	10.13 ± 0.24	10.02 ± 0.50	11.43 ± 0.92	12.42 ± 0.95	18.35 ± 2.36	15.82 ± 1.63
64 / 16	29.50 ± 1.13	16.26 ± 1.02	26.50 ± 10.55	22.07 ± 1.35	20.90 ± 2.25	32.82 ± 4.83	21.59 ± 3.05
32 / 32	42.07 ± 4.16	28.77 ± 2.28	29.94 ± 3.48	35.55 ± 3.12	29.15 ± 0.97	42.97 ± 2.08	45.10 ± 4.46
16 / 64	64.20 ± 6.76	46.06 ± 1.03	50.65 ± 5.66	58.70 ± 7.15	55.40 ± 6.90	57.96 ± 3.65	61.66 ± 4.08

Table 3: Mean and std of the train error versus the size of the hash (k) / size of the reduction (n/k) for the network.

k / $\frac{n}{k}$	Circulant [%]	Random [%]	BinPerm [%]	BinCirc [%]	HalfShift [%]	Toeplitz [%]	VerHorShift [%]
1024 / 1	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.30 ± 0.44	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
512 / 2	0.04 ± 0.06	0.00 ± 0.00	0.00 ± 0.00	2.66 ± 2.98	1.44 ± 2.89	0.00 ± 0.00	0.00 ± 0.01
256 / 4	6.46 ± 2.27	0.00 ± 0.00	0.79 ± 1.57	0.60 ± 1.19	0.49 ± 0.93	2.09 ± 1.69	3.98 ± 3.96
128 / 8	16.89 ± 6.57	4.69 ± 0.43	4.44 ± 0.50	5.62 ± 1.03	7.34 ± 1.27	11.82 ± 2.17	10.51 ± 1.27
64 / 16	26.47 ± 0.98	13.35 ± 0.61	23.98 ± 11.54	18.68 ± 0.78	17.64 ± 2.01	29.97 ± 5.29	18.68 ± 3.26
32 / 32	40.79 ± 3.82	27.51 ± 2.04	28.28 ± 3.23	33.91 ± 3.23	27.90 ± 1.05	41.49 ± 2.14	43.51 ± 3.78
16 / 64	63.96 ± 5.62	46.31 ± 0.73	50.03 ± 6.18	58.71 ± 6.96	54.88 ± 6.47	57.72 ± 3.42	60.91 ± 4.53

and

$$\alpha_{y,i} = \sum_{j=1}^{n} y_j^2 y_{j+i}^2, \tag{39}$$

where indices are added modulo n and this time we cannot assume that all $|x_i|, |y_i|$ are small. Fortunately we have:

$$\sum_{i=1}^{n} \alpha_{x,i} = 1, \tag{40}$$

and

$$\sum_{i=1}^{n} \alpha_{y,i} = 1 \tag{41}$$

Let us fix some positive function $f(k)$. We can conclude that the number of variables $\alpha_{x,i}$ such that $\alpha_{x,i} \geq \frac{f(k)}{\binom{k}{2}}$ is at most $\frac{\binom{k}{2}}{f(k)}$. Note that each such $\alpha_{x,i}$ and each such $\alpha_{y,i}$ corresponds to a pair $\{i_1, i_2\}$ of rows of the matrix $\mathcal{P}$ and consequently to the unique element $Cov(U_{i_1}, U_{i_2})$ of the entire covariance sum (scaled by $\frac{1}{k^2}$). Since trivially we have $|Cov(U_{i_1}, U_{i_2})| = O(1)$, we conclude that the contribution of these elements to the entire covariance sum is of order $\frac{1}{f(k)}$. Let us now consider these $\alpha_{x,i}$ and $\alpha_{y,i}$ that are at most $\frac{f(k)}{\binom{k}{2}}$. These sums are small (if we take $f(k) = o(k^2)$) and thus it makes sense to apply Lemma 7.2 to them. That gives us upper bound $a = \Omega(\Delta)$ with probability:

$$\mathbb{P}^* \geq 1 - e^{-\Omega(a^2 \frac{k^2}{f(k)})}. \tag{42}$$

Taking $f(k) = (\frac{k^2}{\log(k)})^{\frac{1}{3}}$ and $a = O(\Delta) = \frac{1}{f(k)}$, we get: $\mathbb{P}^* \geq 1 - O(\frac{1}{k})$ and furthermore:

$$Var(\tilde{\theta}_{p,r}^n) = \frac{1}{k} \frac{\theta_{p,r}(\pi - \theta_{p,r})}{\pi^2} + (\frac{\log(k)}{k^2})^{\frac{1}{3}}. \tag{43}$$

Thus, from the Chebyshev's inequality, we get the following for every $c > 0$ and fixed points p, r:

$$\mathbb{P}(|\tilde{\theta}_{p,r}^n - \frac{\theta_{p,r}}{\pi}| \geq c(\frac{\sqrt{\log(k)}}{k})^{\frac{1}{3}}) = O(\frac{1}{c^2}). \tag{44}$$

That completes the proof of Theorem 4.2.

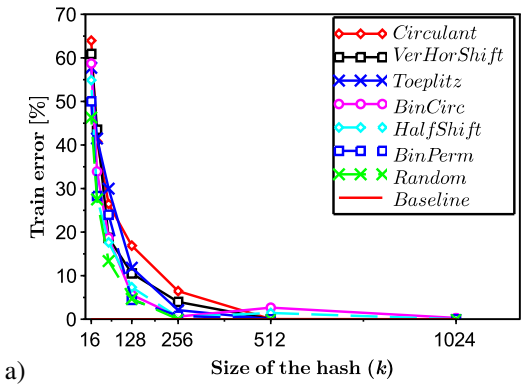
9. Additional figures

Figure 8a and Figure 8b show how the mean train error is affected by the size of the hash, and Figure 8c shows how the mean train error changes with the size of the reduction for the neural network experiment. In Table 3 we report both the mean and the standard deviation of the train error across our neural network experiments. *Baseline* refers to the network with one hidden layer containing 100 hidden units, where all parameters are trained.

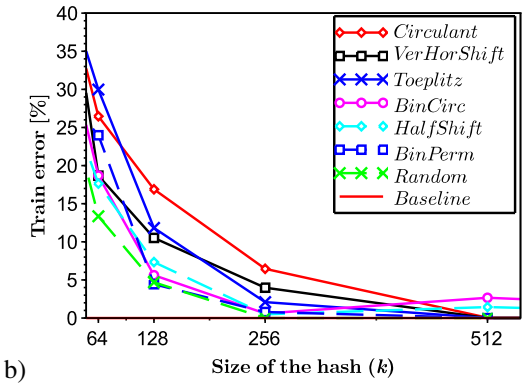
Figure 9a shows the original version of Figure 6a (before zoom). Figure 9b shows the original version of Figure 7a (before zoom). Finally, Table 4 shows the mean and the standard deviation of the test error versus the size of the hash (k)/size of the reduction (n/k) for 1-NN.

Table 4: Mean and std of the test error versus the size of the hash (k) / size of the reduction (n/k) for 1-NN.

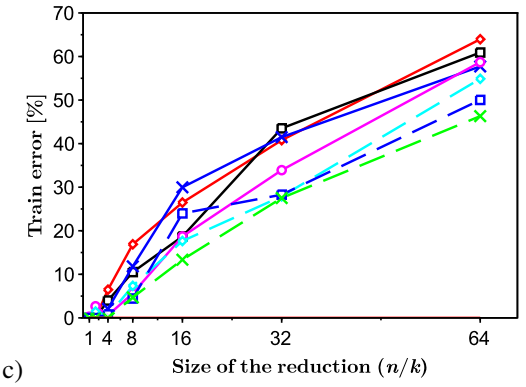
k / $\frac{n}{k}$	Circulant [%]	Random [%]	BinPerm [%]	BinCirc [%]	HalfShift [%]	Toeplitz [%]	VerHorShift [%]
1024 / 1	6.02 ± 0.64	4.83 ± 0.19	6.67 ± 0.65	12.77 ± 2.86	6.38 ± 0.44	6.22 ± 1.20	6.30 ± 0.76
512 / 2	12.98 ± 11.29	5.77 ± 0.11	8.15 ± 0.56	12.40 ± 2.32	7.25 ± 0.71	9.11 ± 2.28	10.81 ± 4.31
256 / 4	17.73 ± 6.66	8.51 ± 0.35	11.11 ± 1.15	12.13 ± 4.35	12.05 ± 2.94	15.66 ± 3.36	18.19 ± 5.46
128 / 8	34.80 ± 14.59	14.44 ± 0.89	17.20 ± 2.26	22.15 ± 6.45	24.74 ± 8.14	33.90 ± 13.90	30.37 ± 7.52
64 / 16	45.91 ± 5.50	27.57 ± 1.58	29.53 ± 3.40	35.33 ± 5.58	36.58 ± 10.71	51.10 ± 13.98	41.66 ± 8.08
32 / 32	65.06 ± 9.60	40.58 ± 2.49	43.58 ± 4.66	53.05 ± 5.39	47.18 ± 7.19	58.24 ± 8.87	56.73 ± 6.09
16 / 64	68.61 ± 5.72	58.72 ± 3.08	60.30 ± 6.11	66.29 ± 4.79	60.84 ± 5.31	72.50 ± 6.04	72.50 ± 5.91



a)

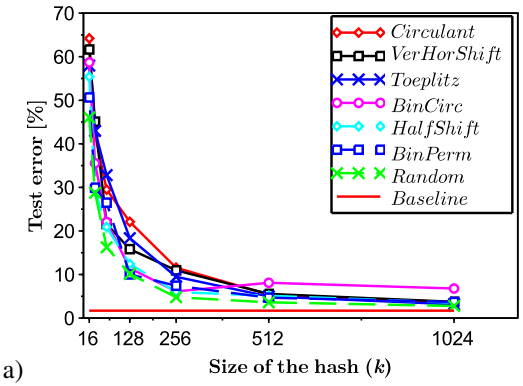


b)

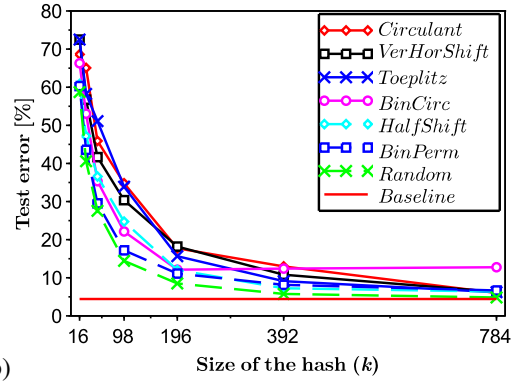


c)

Figure 8: Mean train error versus a), b) the size of the hash (k), c) the size of the reduction (n/k) for the network. b) is a zoomed a). Baseline corresponds to 0%.



a)



b)

Figure 9: Mean test error versus the size of the hash (k) (original plot) for a) the network, b) 1-NN.