Efficient data hashing with structured binary embeddings

Krzysztof Choromanski
Google Research

Abstract. We present here new mechanisms for hashing data via binary embeddings. Contrary to most of the techniques presented before, the embedding matrix of our mechanism is highly structured. That enables us to perform hashing more efficiently and use less memory. What is crucial and nonintuitive is the fact that imposing structured mechanism does not affect the quality of the produced hash. To the best of our knowledge, we are the first to give strong theoretical guarantees of the proposed binary hashing method by proving the efficiency of the mechanism for several classes of structured projection matrices. As a corollary, we obtain binary hashing mechanisms with strong concentration results for circulant and Toeplitz matrices. Our approach is however much more general.

1 Hashing mechanism

In this section we explain in detail proposed hashing mechanism for initial dimensionality reduction that is used to preprocess data before it is given as an input to the autoencoder. As mentioned earlier, the mechanism is of its own interest. We introduce first the aforementioned family of $\Psi$-regular matrices $P$ that is a key ingredient of the method.

Assume that $k$ is the size of the hash and $n$ is the dimensionality of the data. Let $t$ be the size of the pool of independent random gaussian variables $\{g_1, ..., g_t\}$, where each $g_i \sim \mathcal{N}(0, 1)$. Assume that $k \leq n \leq t \leq kn$. We say that a random matrix $P$ is $\Psi$-regular if $P$ is of the form:

$$
\begin{pmatrix}
\sum_{l \in S_{1,j}} g_l & \cdots & \sum_{l \in S_{1,n}} g_l \\
\vdots & \ddots & \vdots \\
\sum_{l \in S_{i,j}} g_l & \cdots & \sum_{l \in S_{i,n}} g_l \\
\sum_{l \in S_{k,j}} g_l & \cdots & \sum_{l \in S_{k,n}} g_l
\end{pmatrix}
$$

(1)

where $S_{i,j} \subseteq \{1, ..., t\}$ for $i \in \{1, ..., k\}$, $j \in \{1, ..., n\}$, $|S_{i,j}| = ... = |S_{i,n}|$ for $i = 1, ..., k$, $S_{i,j} \cap S_{i,u} = \emptyset$ for $i \in \{1, ..., k\}$, $\{j, u\} \subseteq \{1, ..., n\}$, $j \neq u$ and furthermore the following holds:

- for every column of $P$ every $g_l$ appears in at most $\Phi$ entries from that column.

Notice that all structured matrices that we mentioned in the abstract are special cases of the 0-regular matrix. Indeed, each Toeplitz matrix is clearly 0-regular, where subsets $S_{i,j}$ are singletons.
Let \( \phi \) be a function satisfying \( \lim_{x \to \infty} \phi(x) = 1 \) and \( \lim_{x \to -\infty} \phi(x) = -1 \). We will consider two hashing methods. The first one, called by us \textit{extended \( \Psi \)-regular hashing}, applies first random diagonal matrix \( R \) to the datapoint \( x \), then the \( L_2 \)-normalized Hadamard matrix \( H \), next another random diagonal matrix \( D \), then the \( \Psi \)-regular projection matrix \( P_\Psi \) and finally function \( \phi \) (the latter one applied pointwise). The overall scheme is presented below:

\[
x \xrightarrow{R} x_R \xrightarrow{H} x_H \xrightarrow{D} x_D \xrightarrow{P_\Psi} x_{P_\Psi} \xrightarrow{\phi} h(x) \in \mathbb{R}^k.
\] (2)

The diagonal entries of matrices \( R \) and \( D \) are chosen independently from the binary set \( \{-1, 1\} \), each value being chosen with probability \( \frac{1}{2} \). We also propose a shorter pipeline, called by us \textit{short \( \Psi \)-regular hashing}, where we avoid applying first random matrix and Hadamard matrix \( R \) and the Hadamard matrix, i.e. the overall pipeline is of the form:

\[
x \xrightarrow{D} x_D \xrightarrow{P_\Psi} x_{P_\Psi} \xrightarrow{\phi} h(x) \in \mathbb{R}^k.
\] (3)

The goal is to compute good approximation of the angular distance between given \( L_2 \)-normalized vectors \( p, r \), given their compact hashed versions: \( h(p), h(r) \). To achieve this goal we consider the \( L_1 \)-distance in the \( k \)-dimensional space of hashes. Let \( \theta_{p,r} \) denote the angle between vectors \( p \) and \( r \). We define the \textit{normalized approximate angle} between \( p \) and \( r \) as:

\[
\tilde{\theta}_{p,r}^\Pi = \frac{1}{2k} \|h(p) - h(r)\|_1.
\] (4)

In the next section we will show that the normalized approximate angle between vectors \( p \) and \( r \) is a very precise estimation of the actual angle if the chosen parameter \( \Psi \) is not large enough. Furthermore, we show an intriguing connection between theoretical guarantees regarding the quality of the produced hash and the chromatic number of some specific undirected graph encoding the structure of \( P \). For many of the structured matrices under consideration this graph is induced by an algebraic group operation defining the structure of \( P \) (for instance, for the circular matrix the group is a single shift and the underlying graph is a collection of pairwise disjoint cycles and trees thus its chromatic number is at most 3).

2 Theoretical results

2.1 Introduction

We are ready to provide theoretical guarantees regarding the quality of the produced hash. Our guarantees will be given for a \textit{sign} function, i.e for \( \phi \) defined as: \( \phi(x) = 1 \) for \( x \geq 0 \), \( \phi(x) = -1 \) for \( x < 0 \). However we should emphasize that empirical results showed that other functions (that are often used as nonlinear maps in deep neural networks) such as sigmoid function, also work well. It is not hard to show that \( \tilde{\theta}_{p,r}^\Pi \) is an unbiased estimator of \( \theta_{p,r}^\Pi \), i.e. \( E(\tilde{\theta}_{p,r}^\Pi) = \theta_{p,r}^\Pi \).
What we will focus on is the concentration of the random variable \( \tilde{\theta}_{p,r}^n \) around its mean \( \frac{\theta}{\Pi} \). We will prove strong exponential concentration results regarding the extended \( \Psi \)-regular hashing method. Interestingly, the application of the Hadamard mechanism is not necessary and it is possible to get concentration results, yet weaker than in the former case, also for short \( \Psi \)-regular hashing. As a warm up, let us prove the following.

**Lemma 1.** Let \( \mathcal{M} \) be a \( \Psi \)-regular hashing model (either extended or short). Then \( \tilde{\theta}_{p,r}^n \) is an unbiased estimator of \( \theta_{p,r} \), i.e.

\[
E(\tilde{\theta}_{p,r}^n) = \frac{\theta_{p,r}}{\Pi}
\]

**Proof.** Notice first that the \( i \)th row, call it \( g_i \), of the matrix \( P \) is an \( n \)-dimensional gaussian vector with mean 0 and where each element has standard deviation \( \sigma_i \) for \( \sigma_i = |S_{1,i}| = ... = |S_{n,i}| \) (\( i = 1, ..., k \)). Thus, after applying matrix \( D \) the new vector \( g_i \) is still gaussian and of the same distribution. Let us consider first the short \( \Psi \)-regular hashing model. Fix some \( L_2 \)-normalized vectors \( p, r \) (without loss of generality we may assume that they are not collinear) and denote by \( H_{p,r} \) the 2-dimensional hyperplane spanned by \{p, r\}. Denote by \( g_i \) the projection of \( g_i \) into \( H \) and by \( g_i \) the line in \( H \) perpendicular to \( g_i \). Let \( \phi \) be a sign function. Notice that the contribution to the \( L_1 \)-sum \( ||h(p) - h(r)||_1 \) comes from those \( g_i \) for which \( g_i \) divides an angle between \( p \) and \( r \), i.e. from those \( g_i \) for which \( g_i \) is inside the union \( U_{p,r} \) of two 2-dimensional cones bounded by two lines in \( H \) perpendicular to \( p \) and \( r \) respectively. Observe that, from what we have just said, we can conclude that

\[
\tilde{\theta}_{p,r}^n = \sum_{i=1}^{k} X_i \frac{\theta_{p,r}}{\Pi},
\]

where:

\[
X_i = \begin{cases} 
1 & \text{if } g_i \in U_{p,r}, \\
0 & \text{otherwise.}
\end{cases}
\]

Now it suffices to notice that vector \( g_i \) is a gaussian random variable and thus its direction is uniformly distributed over all directions. Thus each \( X_i \) is nonzero with probability exactly \( \frac{\theta}{\Pi} \) and the theorem follows. For the extended \( \Psi \)-regular hashing model the analysis is very similar. The only difference is that data is preprocessed by applying \( HR \) linear mapping first. Both \( H \) and \( R \) are matrices of rotations though, thus their product is also a rotation matrix. Since rotations do not change angular distance, the former analysis can be applied again and yields the proof.

### 2.2 The \( \mathcal{P} \)-chromatic number

As we have already mentioned, the highly well organized structure of the projection matrix \( \mathcal{P} \) gives rise to the underlying undirected graph that encodes dependencies between different entries of \( \mathcal{P} \). More formally, let us fix two rows of \( \mathcal{P} \) of indices \( 1 \leq k_1 < k_2 \leq k \). We define a graph \( \mathcal{G}_\mathcal{P}(k_1, k_2) \) as follows:

\[
V(\mathcal{G}_\mathcal{P}(k_1, k_2)) = \{ j_1, j_2 \} : \exists l \in \{1, ..., t\} \text{s.t. } g_l \in S_{k_1,j_1} \cap S_{k_2,j_2}, j_1 \neq j_2,
\]

where \( S_{i,j} \) is another set of indices.
there exists an edge between vertices \{j_1, j_2\} and \{j_3, j_4\} iff \{j_1, j_2\} \cap \{j_3, j_4\} \neq \emptyset.

The chromatic number \(\chi(G)\) of the graph \(G\) is the minimal number of colors that can be used to color the vertices of the graph in such a way that no two adjacent vertices have the same color.

**Definition 1.** Let \(P\) be a \(Ψ\)-regular matrix. We define the \(P\)-chromatic number \(χ(P)\) as:

\[
χ(P) = \max_{1 \leq k_1, k_2 \leq k} χ(G(k_1, k_2)).
\]

### 2.3 Concentration inequalities for structured hashing with sign function

We present now our main theoretical results. Let us consider first the extended \(Ψ\)-regular hashing model. The following is true:

\[
Ψ(M) \leq Ω(\sqrt{N}k^2). \quad \text{and} \quad μ = \frac{8k(\alpha(Ψ) + \psi(2\epsilon))}{θ_{p,r}}.
\]

Notice how the upper bound on the probability of failure \(P_e\) depends on the \(P\)-chromatic number. The theorem above guarantees strong concentration of \(θ_{p,r}\), around its mean and therefore justifies theoretically the effectiveness of the structured hashing method. It becomes more clearly below.

As a corollary, we obtain the following result:

**Theorem 2.** Take the extended \(Ψ\)-regular hashing model \(M\) with \(t\) independent gaussian random variables: \(g_1, \ldots, g_t\), each of distribution \(N(0,1)\). Let \(N\) be the size of the dataset. Denote by \(k\) the size of the hash and by \(n\) the dimensionality of the data. Let \(f(n)\) be arbitrary positive function. Let \(p, r\) be two fixed vectors \(p, r \in \mathbb{R}^n\) with angular distance \(θ_{p,r}\) between them. Then for every \(a, \epsilon > 0\) the following is true:

\[
P(|θ_{p,r} - k \frac{1}{n}| \leq \epsilon) \geq (1 - 4\left(\frac{N}{2}\right)e^{\frac{r^2(n)}{4}}) e^{-2\epsilon^2}\left(1 - \frac{3\epsilon}{2}\right) = \frac{8k(\alpha(Ψ) + \psi(2\epsilon))}{θ_{p,r}}. \quad \text{and} \quad μ = \frac{8k(\alpha(Ψ) + \psi(2\epsilon))}{θ_{p,r}}.
\]

Theorem 2 follows from Theorem 1 by taking: \(a = n^{-\frac{1}{2}}, \epsilon = k^{-\frac{1}{2}}, \psi(n) = 3\sqrt{\log(n)}\) and noticing that every Toeplitz matrix is 0-regular and the corresponding \(P\)-chromatic number \(χ(P)\) is at most 3.

Let us switch now to the short \(Ψ\)-regular hashing model. The theorem presented below is the application of the Chebyshev’s inequality preceded by the careful analysis of the variance \(Var(θ_{p,r})\).
Theorem 3. Take the short $\Psi$-regular hashing model $M$, where $P$ is a Toeplitz matrix. Let $N$ be the size of the dataset. Denote by $k$ the size of the hash and by $n$ the dimensionality of the data. Let $p, r$ be two vectors $p, r \in \mathbb{R}^n$ with angular distance $\theta_{p, r}$ between them. Then the following is true for any $c > 0$:

$$
P(\|\hat{\theta}_{p, r}^n - \theta_{p, r}^n\| \geq c \frac{\sqrt{\log(k)}}{k}) = O\left(\frac{1}{c^2}\right).
$$

The proofs of Theorem 1 and Theorem 3 will be given in the Appendix.

3 Appendix

In this section we prove Theorem 1 and Theorem 3. We will use notation from Lemma 1.

3.1 Proof of Theorem 1

We start with the following technical lemma:

Lemma 2. Let $\{Z_1, ..., Z_k\}$ be the set of $k$ independent random variables defined on $\Omega$ such that each $Z_i$ has the same distribution and $Z_i \in \{0, 1\}$. Let $\{F_1, ..., F_k\}$ be the set of events, where each $F_i$ is in the $\sigma$-field defined by $Z_i$ (in particular $F_i$ does not depend on the $\sigma$ field $\sigma(Z_1, ..., Z_{i-1}, Z_{i+1}, ..., Z_k)$). Assume that there exists $\mu < \frac{1}{2}$ such that: $P(F_i) \leq \mu$ for $i = 1, ..., k$. Let $\{U_1, ..., U_k\}$ be the set of $k$ random variables such that $U_i \in \{0, 1\}$ and $U_i|F_i = Z_i|F_i$ 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 $Y = \frac{Y_1 + ... + Y_k}{k}$. Then the following is true.

$$
P(\|Y - EY\| > a) \leq \frac{1}{H} \sum_{r=0}^{k} \frac{1}{\sqrt{r}} \frac{k^r}{r^r} \mu^r (1 - \mu)^{k-r} + 2e^{-\frac{a^2}{2k}}. \quad (6)
$$

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

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

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

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

Notice that $P(\mathcal{F}_Q) \leq \mu^r (1 - \mu)^{k-r}$. It follows directly from the Bernoulli scheme.
Denote \( X = \frac{X_1 + \ldots + X_k}{k} \). From what we have just said and from the definition of \( \{F_1, \ldots, F_k\} \) we conclude that for any given \( c \) the following holds:

\[
P(|Y - X| > c) \leq \sum_{r=0}^{k} \binom{k}{r} \mu^r (1 - \mu)^{k-r}.
\]  

(9)

Notice also that from the assumptions of the lemma we trivially get:

\[
E(Y) = E(X).
\]

Let us consider now the expression

\[
P(|Y - E(Y)| > a) = \sum_{r=0}^{k} \binom{k}{r} \mu^r (1 - \mu)^{k-r}.
\]  

(10)

We get:

\[
P(|Y - E(Y)| > a) = \sum_{r=0}^{k} \binom{k}{r} \mu^r (1 - \mu)^{k-r} \leq \sum_{r=0}^{k} \binom{k}{r} \mu^r (1 - \mu)^{k-r} \]

(12)

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

\[
\xi \leq (1 + o_r(1)) \sum_{r=0}^{k} \frac{k^r e^r}{2\pi r^{3/2}} \mu^r (1 - \mu)^{k-r} \leq \frac{1}{2^k} \sum_{r=0}^{k} \frac{k^r e^r}{r^{3/2}} \mu^r (1 - \mu)^{k-r}
\]  

(13)

for \( r \) large enough.

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

**Lemma 3.** Let \( W_1, \ldots, W_k \) be \( k \) independent random variables such that \( E(W_1) = \ldots E(W_k) = 0 \). Assume that \( -\alpha_i \leq W_i+1 - W_i \leq \beta_i \) for \( i = 2, \ldots, k - 1 \). Then the following is true:

\[
P\left( \sum_{i=1}^{k} W_i > a \right) \leq 2e^{-\frac{a^2}{2\sum_{i=1}^{k} \alpha_i \beta_i}}.
\]

Now, using Lemma 3 for \( W_i = X_i - E(X_i) \) and \( \alpha_i = E(X_i), \beta_i = 1 - E(X_i) \) we obtain:

\[
P(|X - E(X)| > a) \leq 2e^{-\frac{a^2}{2}}.
\]  

(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 4.** Let $n$ denote data dimensionality and let $f(n)$ be an arbitrary positive function. Let $D$ be the set of all $L_2$-normalized datapoints, where no two datapoints 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 $HR$, each hyperplane $H_{p,r}$ is transformed into another hyperplane $H^{\Phi R}_{p,r}$. Furthermore, the probability $P_{HR}$ that for every $H^{\Phi R}_{p,r}$ there exist two orthonormal vectors $x = (x_1, ..., x_n), y = (y_1, ..., y_n)$ in $H^{\Phi R}_{p,r}$ such that:

$$|x_i|, |y_i| \leq \frac{f(n)}{\sqrt{n}}$$

satisfies:

$$P_{HR} \geq 1 - 4 \left(\frac{N}{2}\right) e^{-\frac{f^2(n)}{2n}}.$$  

**Proof.** We have already noticed in the proof of Lemma 1 that $HR$ is a matrix of the rotation transformation. 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 $u^{HR}$ vector obtained from $u$ after applying transformation $HR$. Notice that the $j^{th}$ coordinate of $u^{HR}$ is of the form:

$$u^{HR}_j = u_1 T_1 + ... + u_n T_n, \quad (15)$$

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

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

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 3 and the fact that $||u||_2 = 1$, we get for any $a > 0$:

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

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

To complete the proof, it suffices to take $a = f(n)$ and apply the union bound over all vectors $u^{HR}, v^{HR}$ 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 $E_f$. Notice that whether $E_f$ holds or not is determined only by $H, R$ and the initial dataset $D$.

Let us proceed with the proof of Theorem 1. Let us assume that event $E_f$ 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 $f$. 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 $i$th row of $P$ and denote it as $(p_{1,i}, ..., p_{n,i})$. After being multiplied by the diagonal matrix $D$ we obtain another vector:

$$w = (P_{1,i}d_1, ..., P_{n,i}d_n),$$  \hspace{1cm} (18)

where:

$$D_{i,j} = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix}. \hspace{1cm} (19)$$

We have already noticed that in the proof of Lemma 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 $U_{p,r}$. Let us write down the coordinates of the projection of $w$ with $x$ and $y$ respectively thus in the $\{x, y\}$-coordinate system we can write $w$ as:

$$w_{\{x,y\}} = (P_{1,i}d_1x_1, ..., P_{n,i}d_nx_n, P_{1,i}d_1y_1, ..., P_{n,i}d_ny_n). \hspace{1cm} (20)$$

Notice 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 notice that from our assumption about the structure of $P$ we can conclude that both coordinates may be represented as sums of weighted gaussian random variables $g_i$ for $i = 1, ..., t$, i.e.:

$$w_{\{x,y\}} = (g_1s_{i,1} + ... + g_t s_{i,t}, g_1v_{i,1} + ... + g_t v_{i,t}), \hspace{1cm} (21)$$

where each $s_{i,j}, v_{i,j}$ is of the form $d_xz$ or $d_yz$ for some $z$ that depends only on $i, j$. Notice also that

$$s_{i,1}^2 + ... + s_{i,t}^2 = v_{i,1}^2 + ... + v_{i,t}^2. \hspace{1cm} (22)$$

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 5.** 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 $\mathbb{P}_{\text{succ}} \geq 1 - 4(k^2e)\frac{2\log n}{n}$, taken under coin tosses used to construct $D$, the following is true for every $1 \leq i_1 \neq i_2 \leq k$:

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

**Proof.** Notice that the we get the first inequality for free from the fact that \(x\) is orthogonal to \(y\) (in other words, \(\sum_{u=1}^{n} s_{i,u} v_{i,1,u}\) 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. Notice that they can be rewritten as:

\[
E = \sum_{i=1}^{n} d_{\rho(i)} d_{\lambda(i)} x_{\zeta(i)} x_{\gamma(i)}
\]  

(23)

or

\[
E = \sum_{i=1}^{n} d_{\rho(i)} d_{\lambda(i)} y_{\zeta(i)} y_{\gamma(i)}
\]  

(24)

or

\[
E = \sum_{i=1}^{n} d_{\rho(i)} d_{\lambda(i)} x_{\zeta(i)} y_{\gamma(i)}
\]  

(25)

for some \(\rho, \lambda, \zeta, \gamma\). Notice also that from the \(\Psi\)-regularity condition we immediately obtain that \(\rho(i) = \lambda(i)\) for at most \(\Psi\) elements of each sum. Get rid of these elements from each sum and consider the remaining ones. From the definition of the \(P\)-chromatic number, those remaining ones can be partitioned into at most \(\chi(P)\) 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 3. Thus one can conclude that the sum differs from its expectation (which clearly is zero since \(E(d_{i} d_{j}) = 0\) for \(i \neq j\)) by a with probability at most

\[
P_{a} \leq 2e^{-\sum_{i=1}^{n} \frac{\rho^2}{\zeta(i) \gamma(i)}}
\]  

(26)

or

\[
P_{a} \leq 2e^{-\sum_{i=1}^{n} \frac{\rho^2}{\zeta(i) \gamma(i)}}
\]  

(27)

or

\[
P_{a} \leq 2e^{-\sum_{i=1}^{n} \frac{\rho^2}{\zeta(i) \gamma(i)}}
\]  

(28)

Now it is time to use the fact that event \(E_f\) 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 3 and then taking the union bound, we complete the proof.
We can finish the proof of Theorem 1. From Lemma 5 we see that $s_1, \ldots, s_k, v_1, \ldots, v_k$ are close to pairwise orthogonal with high probability. Let us fix some positive function $f(n) > 0$ and some $a > 0$. Denote

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

Notice that, by Lemma 5 we see that applying Gram-Schmidt process we can obtain a system of pairwise orthogonal vectors $\tilde{s}_1, \ldots, \tilde{s}_k, \tilde{v}_1, \ldots, \tilde{v}_k$ such that

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

and

$$\|\tilde{s}_i - s_i\|_2 \leq k \Delta. \quad (31)$$

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 $\mathcal{P}$. Notice 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 $\tilde{s}_1, \ldots, \tilde{s}_k, \tilde{v}_1, \ldots, \tilde{v}_k$ are pairwise orthogonal. Notice also that from (30) and (31) we obtain that the angular distance between $w_{x,y}$ and $\tilde{w}_{x,y}$ is at most $k \Delta$.

Let $Z_i$ for $i = 1, \ldots, k$ be an indicator random variable that is zero if $\tilde{w}_{x,y}$ is inside the region $U_{p,r}$ and zero otherwise. Let $U_i$ for $i = 1, \ldots, k$ be an indicator random variable that is zero if $w_{x,y}$ is inside the region $U_{p,r}$ and zero otherwise. Notice that $\tilde{\theta}_{p,r} = \frac{\theta_{p,r}}{k \Delta}$, $\tilde{\theta}_{p,r} = \frac{\theta_{p,r}}{k \Delta}$. Furthermore, random variables $Z_1, \ldots, Z_k, U_1, \ldots, U_k$ satisfy the assumptions of Lemma 2 with $\mu \leq \frac{\theta_{p,r}}{k \Delta}$. 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 $\frac{\theta_{p,r}}{k \Delta}$. Furthermore $Z_i \neq U_i$ only if $w_{x,y}$ is inside $U_{p,r}$ and $\tilde{w}_{x,y}$ is outside $U_{p,r}$, 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 $U_{p,r}$, namely within an angular distance $\frac{\theta_{p,r}}{k \Delta}$ from one of the four semilines defining $U_{p,r}$. Thus in particular an event $Z_i \neq U_i$ is contained in the event of probability at most $2 \cdot 4 \cdot \frac{\theta_{p,r}}{k \Delta}$ that depends only on one $w_{x,y}$.

But then we can apply Lemma 2. All we need is to assume that the premises of Lemma 2 are satisfied. But this is the case with probability specified in Lemma 4 and this probability is taken under random coin tosses used to produce $\mathcal{H}$ and $\mathcal{R}$, thus independently from the random coin tosses used to produce $D$. Putting it all together we obtain the statement of Theorem 1.

### 3.2 Proof of Theorem 3

We will borrow some notation from the proof of Theorem 1. Notice however that in this setting no preprocessing with the use of matrices $\mathcal{H}$ and $\mathcal{R}$ is applied.

**Lemma 6.** Define $U_1, \ldots, U_k$ as in the proof of Theorem 1. Assume that the following is true:

$$\left| \sum_{u=1}^{n} s_{i,u} v_{i,u} \right| \leq \Delta.$$
\[ \left| \sum_{u=1}^{n} s_{i_1,u} s_{i_2,u} \right| \leq \Delta, \]
\[ \left| \sum_{u=1}^{n} v_{i_1,u} v_{i_2,u} \right| \leq \Delta, \]
\[ \left| \sum_{u=1}^{n} s_{i_1,u} v_{i_2,u} \right| \leq \Delta. \]

for some \( 0 < \Delta < 1 \). 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 1, thus we leave it to the reader as an exercise.

Notice that we have:
\[ \text{Var}(\tilde{\theta}^n_{p,r}) = \text{Var}\left( \frac{U_1 + \cdots + U_k}{k} \right) = \frac{1}{k^2} \left( \sum_{i=1}^{k} \text{Var}(U_i) + \sum_{i \neq j} \text{Cov}(U_i, U_j) \right). \tag{32} \]

Since \( U_i \) is an indicator random variable that takes value one with probability \( \frac{\theta}{\Pi} \), we get:
\[ \text{Var}(U_i) = E(U_i^2) - E(U_i)^2 = \frac{\theta}{\Pi} (1 - \frac{\theta}{\Pi}). \tag{33} \]

Thus we have:
\[ \text{Var}(\tilde{\theta}^n_{p,r}) = \frac{\theta}{k} \left( \frac{\Pi - \theta}{\Pi^2} \right) + \frac{1}{k^2} \sum_{i \neq j} \text{Cov}(U_i, U_j). \tag{34} \]

Notice however that \( \text{Cov}(U_i, U_j) \) is exactly: \( \mathbb{P}(U_i U_j = 1) - \mathbb{P}(U_i = 1)\mathbb{P}(U_j = 1) \). Therefore, using Lemma 6, we obtain:
\[ \text{Var}(\tilde{\theta}^n_{p,r}) = \frac{\theta}{k} \left( \frac{\Pi - \theta}{\Pi^2} \right) + O(\Delta). \tag{35} \]

It suffices to estimate parameter \( \Delta \). 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 \( P \). Notice that, again, we see that applying Gram-Schmidt process we can obtain a system of pairwise orthogonal vectors \( \tilde{s}_{i_1}, \tilde{s}_{i_2}, \tilde{v}_{i_1}, \tilde{v}_{i_2} \) such that
\[ \| \tilde{v}_{i_1} - v_{i_2} \|_2 \leq \Delta. \tag{36} \]
and
\[ \| \tilde{s}_{i_1} - s_{i_2} \|_2 \leq \Delta. \tag{37} \]
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 $\chi(P)$ subchunks, where this time $\chi(P) \leq 3$ (since $P$ is Toeplitz). The problem is that applying Lemma 3, we get bounds that depend on the expressions of the form

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

and

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

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

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

and

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

Let us fix some positive function $f(k)$. We can conclude that the number of variables $\alpha_{x,i}$ such that $\alpha_{x,i} \geq f(k) \frac{\binom{n}{2}}{f(k)}$ is at most $\frac{\binom{n}{2}}{f(k)}$. Notice that each such $\alpha_{x,i}$ and each such $\alpha_{y,i}$ corresponds to a pair $\{i_{1,2}\}$ of rows of the matrix $P$ and consequently to the unique element $\text{Cov}(U_{i_1}, U_{i_2})$ of the entire covariance sum (scaled by $\frac{1}{\chi(P)}$). Since trivially we have $|\text{Cov}(U_{i_1}, U_{i_2})| = O(1)$, we conclude that the contribution of these elements to the entire covariance sum is of order $\frac{1}{\chi(P)}$.

Let us now consider these $\alpha_{x,i}$ and $\alpha_{y,i}$ that are at most $f(k) \frac{\binom{n}{2}}{f(k)}$. These sums are small (if we take $f(k) = o(k^2)$) and thus it makes sense to apply Lemma 3 to them. That gives us upper bound $a = \Delta$ with probability:

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

Taking $f(k) = (\frac{k^2}{\log(k)})^\frac{1}{2}$ and $a = \Delta = \frac{1}{f(k)}$, we conclude that:

$$\text{Var}(\tilde{\theta}^n_{p,r}) \leq \frac{1}{k} \frac{\theta(\Pi - \theta)}{\Pi^2} + \left(\frac{\log(k)}{k^2}\right)^{\frac{1}{2}}. \tag{43}$$

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

$$\mathbb{P}(|\tilde{\theta}^n_{p,r} - \frac{\theta}{\Pi}| \geq c\left(\frac{\sqrt{\log(k)}}{k}\right)^{\frac{1}{2}}) \leq O\left(\frac{1}{c^2}\right).$$

$$\mathbb{P}(|\tilde{\theta}^n_{p,r} - \frac{\theta}{\Pi}| \geq c\left(\frac{\sqrt{\log(k)}}{k}\right)^{\frac{1}{2}}) = O\left(\frac{1}{c^2}\right). \tag{44}$$
That completes the proof.