7. Appendix

We now prove all theoretical results of the paper. We need to introduce some technical denotation.

From now on f denotes one from the following functions: sin, cos, $sign$ or a linear rectifier. We call the set of these functions $\mathcal{F}$. For two vectors v, w we denote by $v \cdot w$ their dot product. We denote by G^i_{struct} for $i = 1, ..., \frac{k}{m}$ the building blocks of the structured matrix constructed according to the $\mathcal{P}$-model that are vertically stacked to produce the final structured matrix. Let $v^1, v^2 \in \mathbb{R}^n$ be two datapoints from the preprocessed input-dataset $D_1 H D_0 \mathcal{X}$. Let d be a fixed integer constant. Let $R = \{i_1, ..., i_r\}$ be some r-element subset of the set $\{1, ..., m\}$, where m stands for the number of rows used in the construction of matrices G^i_{struct} (key building blocks of our structured mechanism). Finally, let $\alpha_1, ..., \alpha_r$ be positive integers such that $\alpha_1 + ... + \alpha_r = d$.

Definition 7.1. *For three vectors: $v, w, z \in \mathbb{R}^n$ and a given nonlinear function $f \in \mathcal{F}$ we denote:*

$$\phi(v, w, z) = f(z \cdot v) f(z \cdot w).$$

We will show that for a variety of functions $\Psi : \mathbb{R}^r \rightarrow \mathbb{R}$ the expected value of the expression $T^{G,d}_{v^1,v^2}(\mathcal{R}, \alpha_1, ..., \alpha_r)$ given by the formula:

$$\Psi(\phi_1(v^1, v^2, g^{i_1})^{\alpha_1}, ..., \phi_r(v^1, v^2, g^{i_r})^{\alpha_r}), \quad (12)$$

where $g^1, ..., g^m$ is the set of m gaussian vectors forming gaussian matrix G, each obtained by sampling independently n values from the distribution $\mathcal{N}(0, 1)$ and ϕ_is differ by the choice of nonlinear mapping $f_i \in \mathcal{F}$, can be accurately approximated by its structured version $T^{A,d}_{v^1,v^2}((\mathcal{R}, \alpha_1, ..., \alpha_r)$ which is of the form:

$$\Psi(\phi_1(v^1, v^2, a^{i_1})^{\alpha_1}, ..., \phi_r(v^1, v^2, a^{i_r})^{\alpha_r}), \quad (13)$$

where $a^1, ..., a^m$ are rows of the structured matrix $A = G^i_{struct}$. The importance of $T^{G,d}_{v^1,v^2}(\mathcal{R}, \alpha_1, ..., \alpha_r)$ and $T^{A,d}_{v^1,v^2}(\mathcal{R}, \alpha_1, ..., \alpha_r)$ lies in the fact that d^{th} moments of the random variables approximating considered kernels in the unstructured and structured mechanism can be expressed as weighted sums of the expressions of the form $T^{G,d}_{v^1,v^2}(\alpha_1, ..., \alpha_r)$ and $T^{A,d}_{v^1,v^2}(\alpha_1, ..., \alpha_r)$ respectively if $\Psi(x_1, ..., x_r) = x_1 \cdot ... \cdot x_r$. Thus if $T^{A,d}_{v^1,v^2}(\alpha_1, ..., \alpha_r)$ closely approximates $T^{G,d}_{v^1,v^2}(\alpha_1, ..., \alpha_r)$ then the corresponding moments are similar. That, as we will see soon, implies several theoretical guarantees for the structured method. In particular, this means that the variances are similar. Since in the unstructured setting the variance is of the order $O(\frac{1}{m})$, that will be also the case for the structured setting. This in turn will imply concentration results providing theoretical explanation for the observations from the experimental section that show the quality of the proposed structured setting.

We need to introduce a few definitions.

Definition 7.2. *We denote by Δ^{ξ}_s the supremum of the expression $\|\xi(y_1, ..., y_m) - \xi(y'_1, ..., y'_m)\|$ over all pairs of vectors $(y_1, ..., y_m), (y'_1, ..., y'_m)$ from the domain $\mathcal{D}$ that differ on at most one dimension and by at most s. We say that a function $\xi : \mathbb{R}^m \rightarrow \mathbb{R}$ is M-bounded in the domain $\mathcal{D}$ if $\Delta^{\xi}_{\infty} = M$.*

Note that the value of the function $\phi_i(v^1, v^2, g^i)^{\alpha_i}$ depends only on the projection g^i_{proj} of g^i on the 2-dimensional space spanned by v^1 and v^2. Thus for a given pair v^1, v^2 function ϕ is in fact a function B^{v^1,v^2}_i of this projection.

Definition 7.3. *Define:*

$$p_{\lambda,\epsilon} = \sup_{i, v^1, v^2, \|\zeta|_{\infty} \leq \epsilon} \mathbb{P}[|B^{v^1,v^2}_i(g^i_{proj} + \zeta) - B^{v^1,v^2}_i(g^i_{proj})| > \lambda], \quad (14)$$

where the supremum is taken over all indices $i = 1, ..., m$, all pairs of linearly independent vectors from the domain, all coordinate systems in $span(v^1, v^2)$ and vectors ζ of L_1-norm at most ϵ in some of these coordinate systems.

We will use the following notation: $\sigma_{i,j}(n_1, n_2) = \mathbf{P}^T_{i,n_1} \mathbf{P}_{j,n_2}$. To compress the statements of our theoretical results, we will use also the following notation:

$$\xi(i_i, i_2) = 2\chi(i_1, i_2) \sqrt{\sum_{1 \leq n_1 < n_2 \leq n} (\sigma_{i_1,i_2}(n_1, n_2))^2},$$

We will also denote: $\lambda(i_1, i_2) = \sum_{j=1}^n |\sigma_{i_1,i_2}(j, j)|$ and $\tilde{\lambda}(i_1, i_2) = |\sum_{j=1}^n \sigma_{i_1,i_2}(j, j)|$ for $1 \leq i_1 \leq i_2 \leq m$ (see: 3.1).

Note first that the preprocessing step preserves kernels' values since transformation HD_0 is an isometry and considered kernels are spherically-invariant. We start with Lemma 4.1.

Proof. Note that it suffices to show that for any two given vectors $x, y \in \mathbb{R}^n$ the following holds:

$$\mathbb{E}[f(G^i_{struct}x) \cdot f(G^i_{struct}y)] = \mathbb{E}[f(Gx) \cdot f(Gy)], \quad (15)$$

where G is the unstructured gaussian matrix. Let $g^{i,j}_{struct}$ be the j^{th} row of G^i_{struct} and let g^j be the j^{th} row of G. Note that we have:

$$\mathbb{E}[f(g^{i,j}_{struct} \cdot x) f(g^{i,j}_{struct} \cdot y)] = \mathbb{E}[f(g^j \cdot x) f(g^j \cdot y)]. \quad (16)$$

The latter follows from the fact that $g^{i,j}_{struct}$ has the same distribution as g. To see this note that $g^{i,j}_{struct} = g \cdot P_i$. Thus dimensions of $g^{i,j}_{struct}$ are projections of g onto columns of P_i. Each projection is trivially gaussian from $\mathcal{N}(0,1)$ (that is implied by the fact that each column is normalized). The independence of different dimensions of $g^{i,j}_{struct}$ comes from the observation that different columns are orthogonal. Thus we can use a simple property of gaussian vectors stating that the projections of a gaussian vector on mutually orthogonal directions are independent. The equation 15 implies equation 16 by the linearity of expectation and that completes the proof. □

Now we prove Theorem 4.1. This one is easily implied by a more general result that we state below. We will assume that function Ψ from equations: 12, 13 is M-bounded for some given $M > 0$. We will assume that expected values defining $T^{A,d}$ are not with respect to the random choices determining $P_i s$.

Theorem 7.1. *Let $v^1, v^2 \in \mathbb{R}^n$ be two vectors from a dataset $\mathcal{X}$. Let $\mathcal{R} = \{i_1, ..., i_r\} \in \{1, ..., m\}$ and let $\alpha_1, ..., \alpha_r$ be the set of positive integers such that $\alpha_1 + ... + \alpha_r = d$. Assume that each structured matrix G^i_{struct} consists of m rows and either $\sup_{1 \leq i_1 < i_2 \leq m} \lambda(i_1, i_2) = o(\frac{n}{\log^2(n)})$ if $P_i s$ were constructed deterministically or $sup_{1 \leq i_1 < i_2 \leq m} \mathbb{E}[\tilde{\lambda}(i_1, i_2)] = o(\frac{n}{\log^2(n)})$ if $P_i s$ were constructed randomly. In the latter case assume also that for any $1 \leq i_1 < i_2 \leq m$ and $1 \leq n_1 < n_2 \leq n$ the n_1^{th} column of P_{i_1} is chosen independently from the n_2^{th} column of P_{i_2}. Denote by Ψ_{max} the maximal value of the function Ψ for the datapoints from $\mathcal{X}$. Let $q^d_{v^1,v^2} = |T^{A,d}_{v^1,v^2}(\mathcal{R}, \alpha_1, ..., \alpha_r) - T^{G,d}_{v^1,v^2}(\mathcal{R}, \alpha_1, ..., \alpha_r)|$ denote the absolute value of the difference of the two fixed terms on the weighted sum for the d-moments of the kernel's approximation in the structured $\mathcal{P}$-model setting and the fully unstructured setting. Then for any $\lambda, \epsilon > 0$, $T > 0$, n large enough and $P_i s$ chosen deterministically we have:*

$$q^d_{v^1,v^2} \leq (p_{gen} + p_{struct})\Psi_{max} + \sum_{i=0}^{d} p^i_f (iM + (d-i)\Delta^{\Psi}_{\lambda}),$$

where:

$$p_{gen} = \frac{4r}{\sqrt{2\pi T}} e^{-\frac{T}{2}} + 4ne^{-\frac{\log^2(n)}{8}}, \tag{17}$$

$$p^i_f = \binom{d}{i} (p_{\lambda,\epsilon})^i \tag{18}$$

and

$$p_{struct} = 4 \sum_{i=1}^{m} \chi(i,i) e^{-\frac{1}{2\xi^2(i,i)} \frac{n^2}{\log^6(n)}} + 2 \sum_{1 \leq i_1 \leq i_2 \leq m} \chi(i_1, i_2) e^{-\frac{\epsilon^2 n^{\frac{3}{2}}}{2\xi^2(i_1,i_2) T \log^4(n)}} \tag{19}$$

If $P_i s$ are chosen from the probabilistic model then the above holds with probability at least $1 - p_{wrong}$, where $p_{wrong} = 2 \sum_{i \leq i_1 < i_2 \leq m} e^{-\frac{n^2}{8 \log^6(n) \sum_{j=1}^{n} (\sigma_{i_1,i_2}(j,j))^2}}$.

Proof. Consider the expression

$$q^d_{v_1,v_2} = |T^{A,d}_{v^1,v^2}(\mathcal{R}, \alpha_1, ..., \alpha_r) - T^{G,d}_{v^1,v^2}(\mathcal{R}, \alpha_1, ..., \alpha_r)|.$$

We will use formulas for $T^{G,d}$ and $T^{A,d}$ given by equations: 12 and 13. Without loss of generality we will assume that $A = G^i_{struct} D_1$ i.e. in our theoretical analysis we will make D_1 a part of the structured mechanism and move it away from the preprocessing phase (obviously both ways are equivalent because of the associative property of matrix mutliplication). We have already noted that each argument of the function Ψ from equations: 12 and 13 depends only on the projections of $a^{i_1}, ..., a^{i_r}$ on the 2-dimensional space spanned by v^1 and v^2. Denote these projections as: $a^{i_1}_{proj}$,...,$a^{i_r}_{proj}$ respectively and fix some orthonormal basis $\mathcal{B}$ of this 2-dimensional space. As we will see soon, in the $\mathcal{P}$-model setting the coordinates of $a^i_{proj} s$ in $\mathcal{B}$ can be expressed as $g \cdot s^{i,j}$ for $j = 1, 2$, where g is a vector representing a budget of randomness of the corresponding $\mathcal{P}$-model and $s^{i,j}$s are some vectors from $\mathbb{R}^t$ (parameter t stands for the length of g).

We will show that $s^{i,j}$s, even though not necessarily pairwise orthogonal, are close to be pairwise orthogonal with high probability. Let us assume now that vectors $s^{i,j}$ can be chosen in such a way that each $s^{i,j}$ satisfies: $s^{i,j} = w^{i,j} + \rho(i,j)$, where vectors $w^{i,j}$ are mutually orthogonal, we have $\|s^{i,j}\|_2 = \|w^{i,j}\|_2$ and furthermore $\|\rho(i,j)\|_2 \leq \rho$ for some given $\rho > 0$. We call this property the ρ-orthogonality property. We will later show that the ρ-orthogonality property depends on the random diagonal matrix D_1.

Assume now that the ρ-orthogonality property is satisfied. Denote by $g^{\mathcal{H}}$ the projection of the "budget-of-randomness" vector g onto $2r$-dimensional linear space $\mathcal{H}$ spanned by vectors from $\{s^{i,j}\}$. Note that then the coordinates of $a^i_{proj} s$ in $\mathcal{B}$ can be rewritten as $g \cdot w^{i,j} + \epsilon(i,j)$, where $|\epsilon(i,j)| \leq \epsilon$ and $\epsilon = \|g^{\mathcal{H}}\|_2 \rho$. Thus each ψ_i in the formula from equation 13 can be then expressed as $B^{v^1,v^2}_i (g^i_{proj} + \epsilon(i))$, where $g^i_{proj} s$ stand for the projections onto 2-dimensional linear space spanned by v^1 and v^2 of independent copies of gaussian vectors g^i. Each g^i is of the same distribution as the corresponding structured vector a^i and $\epsilon(i)s$ are vectors with the L_1-norm satisfying $\|\epsilon(i)\| \leq \epsilon$. The independence comes from the fact that variables of the form $g \cdot w^{i,j}$ are independent. That, as in the proof of Lemma 4.1 is implied by the well known fact that dot products of a given gaussian vector with orthogonal vectors are independent. Note that if not the term

$\epsilon(i)$ then the formula for $T^{A,d}$ would collapse to its unstructured counterpart $T^{G,d}$. We will argue that both expressions are still close to each other if $\epsilon(i)$ have small L_1-norm.

Let us fix $\lambda > 0$. Our goal is to count these indices i that satisfy the following: $|\psi_i(v^1, v^2, g^i)^{\alpha^i} - \psi_i(v^1, v^2, g^i)^{\alpha^i}| > \lambda$, where $g^i s$ corresponds to the aforementioned independent counterparts of $a^i s$. We call them *bad indices*. Based on what we have said so far, we can conclude that the latter inequality can be expressed as $|B_i^{v^1,v^2}(g^i_{proj} + \epsilon(i)) - B_i^{v^1,v^2}(g^i_{proj})| > \lambda$. Let us first find the upper bound on the probability of the event that the number of bad indices is j for some fixed $1 \leq j \leq d$. Note that since $g^i s$ are independent, we can use Bernoulli scheme to find that upped bound. Using the definition of $p_{\lambda,\epsilon}$ we obtain an upper bound of the form $p_{upper} \leq \binom{d}{j}(p_{\lambda,\epsilon})^j$. If the number of bad indices is j then by the definition of M and Δ^{Ψ}_{λ} we see that $T^{A,d}$ differs from $T^{G,d}$ by at most $iM + (d-i)\Delta^{\Psi}_{\lambda}$. Summing up over all indices j we get the second term of the upper bound on $q^d_{v^1,v^2}$ from the statement of the theorem.

However the ρ-orthogonality does not have to hold. Note that (by the definition of Ψ_{max}) to finish the proof of the theorem it suffices to show that the probability of ρ-orthogonality not to hold is at most $p_{gen} + p_{struct}$.

Lemma 7.1. *The ρ-orthogonality property holds with probability at least $1 - (p_{gen} + p_{struct})$.*

Proof. We need the following definition.

Definition 7.4. *Let $x = (x_1, ..., x_n)$ be a vector with $\|x\|_2 = 1$. We say that x is θ-balanced if $|x_i| \leq \frac{\theta}{\sqrt{n}}$ for $i = 1, ..., n$.*

For a fixed pair of vectors $v^1, v^2 \in \mathcal{X}$ choose some orthonormal basis $\mathcal{B} = \{x^1, x^2\}$ of the 2-dimensional space spanned by v^1 and v^2. Let $\tilde{x}^1$ and $\tilde{x}^2$ be the images of x^1 and x^2 under transformation HD_0, where H is a Hadamard matrix and D_0 is a random diagonal matrix. We will show now that with high probability $\tilde{x}^1$ and $\tilde{x}^2$ are $\log(n)$-balanced. Indeed, the i^{th} dimension of $\tilde{x}^1$ is of the form: $\tilde{x}^1_i = h_{i,1}x^1_1 + ... + h_{i,n}x^1_n$, where $h_{i,j}$ stands for the entry in the i^{th} row and j^{th} column of a matrix HD_0. We need to find a sharp upper bound on $\mathbb{P}[|h_{i,1}x^1_1 + ... + h_{i,n}x^1_n| \geq a]$ for $a = \frac{\log(n)}{\sqrt{n}}$.

We will use the following concentration inequality, calles *Azuma's inequality*

Lemma 7.2. *Let $X_1, ..., X_n$ be a martingale and assume that $-\alpha_i \leq X_i \leq \beta_i$ for some positive constants $\alpha_1, ..., \alpha_n, \beta_1, ..., \beta_n$. Denote $X = \sum_{i=1}^n X_i$. Then the following is true:*

$$\mathbb{P}[|X - \mathbb{E}[X]| > a] \leq 2e^{-\frac{a^2}{2\sum_{i=1}^n(\alpha_i+\beta_i)^2}}$$

In our case $X_j = h_{i,j}x^1_j$ and $\alpha_i = \beta_i = \frac{1}{\sqrt{n}}$. Applying Azuma's inequality, we obtain the following bound: $\mathbb{P}[|h_{i,1}x^1_1 + ... + h_{i,n}x^1_n| \geq \frac{\log(n)}{\sqrt{n}}] \leq 2e^{-\frac{\log^2(n)}{8}}$. The probability that all n dimensions of $\tilde{x}^1$ and $\tilde{x}^2$ have absolute value at most $\frac{\log(n)}{\sqrt{n}}$ is, by the union bound, at least $p_{balanced} = 1 - 2n \cdot 2e^{-\frac{\log^2(n)}{8}} = 1 - 4ne^{-\frac{\log^2(n)}{8}}$. Thus this a lower bound on the probability that $\tilde{x}^1$ and $\tilde{x}^2$ are $\log(n)$-balanced. We will use this lower bound later. Now note that it does not depend on the particular form of the structured matrix since it is only related to the preprocessing phase, where linear mappings D_0 and H are applied.

For simplicity we will now denote $\hat{x}^1$ and $\hat{x}^2$ simply as x^1 and x^2, knowing these are the original vectors after applying linear transformation HD_0. Let us get back to the projections of $a^i s$ onto 2-dimensional linear space spanned by v^1 and v^2. Note that we have already noticed that $a^i \cdot x^j$ $(j = 1, 2)$ is of the form $g \cdot s^{i,j}$ for some vector $s^{i,j} \in \mathbb{R}^t$, where t is the size of the "budget of randomness" used in the given $\mathcal{P}$-model. From the definition of the $\mathcal{P}$-model we obtain:

$$s^{i,j}_l = d_1 p^i_{l,1} x^j_1 + ... + d_n p^i_{l,n} x^j_n \quad (20)$$

for $l = 1, ..., t$, where $s^{i,j}_l$ stands for the l^{th} dimension of $s^{i,j}$, $p^i_{l,k}$ is the entry in the l^{th} row and k^{th} column of P_i and $d_r s$ are the values on the diagonal of the matrix D_0. As we noted earlier, we want to show that $s^{i,j}s$ are close to be mutually orthogonal. To do it, we will compute dot products $s^{i_1,j_1} \cdot s^{i_2,j_2}$. We will first do it for $i_1 = i_2$. We have:

$$s^{i_1,j_1} \cdot s^{i_1,j_2} = x^{j_1}_1 x^{j_2}_1 \sum_{l=1}^t (p^{i_1}_{l,1})^2 + ... + x^{j_1}_n x^{j_2}_n \sum_{l=1}^t (p^{i_1}_{l,n})^2$$
$$+2 \sum_{1 \leq n_1 < n_2 \leq n} d_{n_1} d_{n_2} x^{j_1}_{n_1} x^{j_2}_{n_2} (\sum_{i=1}^t p^{i_1}_{l,n_1} p^{i_2}_{l,n_2}) \quad (21)$$

Now we take advantage of the normalization property of the matrices P_i and the fact that x^1 is orthogonal to x^2 and conclude that the first term on the RHS of the equation above is equal to 0. Thus we have:

$$s^{i_1,j_1} \cdot s^{i_1,j_2} = 2 \sum_{1 \leq n_1 < n_2 \leq n} d_{n_1} d_{n_2} x^{j_1}_{n_1} x^{j_2}_{n_2} \sigma_{i_1,i_1}(n_1, n_2). \quad (22)$$

Note that if for any fixed P_i any two different columns of P_i are orthogonal then $\sigma_{i_1,i_1}(n_1, n_2) = 0$ and thus

$s^{i_1,j_1} \cdot s^{i_1,j_2} = 0$. This is the case for many structured matrices constructed according to the $\mathcal{P}$-model, for instance circulant, Toeplitz or Hankel matrices.

Let us consider now $s^{i_1,j_1} \cdot s^{i_2,j_2}$ for $i_1 \neq i_2$. By the previous analysis, we obtain:

$$s^{i_1,j_1} \cdot s^{i_2,j_2} = \sigma_{i_1,i_2}(1,1)x_1^{j_1}x_1^{j_2} + ... + \sigma_{i_1,i_2}(n,n)x_n^{j_1}x_n^{j_2} + 2\sum_{1 \leq n_1 < n_2 \leq n} d_{n_1}d_{n_2}x_{n_1}^{j_1}x_{n_2}^{j_2}\sigma_{i_1,i_2}(n_1,n_2). \tag{23}$$

This time in general we cannot get rid of the first term in the RHS expression. This can be done if columns of the same indices in different $P_i s$ are orthogonal. This is in fact again the case for circulant, Toeplitz or Hankel matrices.

Let us now fix some $1 \leq i_1 \leq m$ and $\kappa > 0$. Our goal is to find an upper bound on the following probability: $\mathbb{P}[|s^{i_1,j_1} \cdot s^{i_2,j_2}| > \kappa]$.

We have:

$$\mathbb{P}[|s^{i_1,j_1} \cdot s^{i_2,j_2}| > \kappa] = \mathbb{P}[|\sum_{1 \leq n_1 < n_2 \leq n} d_{n_1}d_{n_2}x_{n_1}^{j_1}x_{n_2}^{j_2}2\sigma_{i_1,i_2}(n_1,n_2)| > \kappa]. \tag{24}$$

For $\{n_1, n_2\}$ such that $n_1 \neq n_2$ and $\sigma_{i_1,i_1}(n_1,n_2) \neq 0$ let us now consider random variables Y_{n_1,n_2} that are defined as follows

$$Y_{n_1,n_2} = 2d_{n_1}d_{n_2}x_{n_1}^{j_1}x_{n_2}^{j_2}\sigma_{i_1,i_1}(n_1,n_2). \tag{25}$$

From the definition of the chromatic number $\chi(i_1, i_1)$ we can deduce that the set of all this random variables can be partitioned into at most $\chi(i_1, i_1)$ subsets such that random variables in each subset are independent. Let us denote these subsets as: $\mathcal{L}_1, ..., \mathcal{L}_r$, where $r \leq \chi(i_1, i_1)$. Note that an event $\{|\sum_{1 \leq n_1 < n_2 \leq n} d_{n_1}d_{n_2}x_{n_1}^{j_1}x_{n_2}^{j_2}2\sigma_{i_1,i_1}(n_1,n_2)| > \kappa\}$ is contained in the sum of the events: $\mathcal{E} = \mathcal{E}_1 \cup ... \cup \mathcal{E}_r$, where each $\mathcal{E}_j$ is defined as follows:

$$\mathcal{E}_j = \{|\sum_{Y \in \mathcal{L}_j} Y| \geq \frac{\kappa}{\chi(i_1, i_1)}\}. \tag{26}$$

Thus, from the union bound we get:

$$\mathbb{P}[\mathcal{E}] \leq \sum_{i=1}^{\chi(i_1,i_1)} \mathbb{P}[\mathcal{E}_i]. \tag{27}$$

Now we can use Azuma's inequality to find an upper bound on $\mathcal{P}[\mathcal{E}_i]$ and we obtain:

$$\mathbb{P}[\mathcal{E}_i] \leq 2e^{-\frac{\frac{\kappa^2}{\chi^2(i_1,i_1)}}{2\sum_{1 \leq n_1 < n_2 \leq n}(2\sigma_{i_1,i_1}(n_1,n_2))^2(x_{n_1}^{j_1})^2(x_{n_2}^{j_2})^2}}. \tag{28}$$

Now, if we assume that the vectors of the orthonormal basis $\mathcal{B}$ are $\log(n)$-balanced, then by the union bound we obtain the following upper bound on the probability $\mathbb{P}[\mathcal{E}]$:

$$\mathbb{P}[\mathcal{E}] \leq 2\chi(i_1, i_1)e^{-\frac{\kappa^2 n^2}{2\log^4(n)\chi^2(i_1,i_1)\sum_{1 \leq n_1 < n_2 \leq n}(2\sigma_{i_1,i_1}(n_1,n_2))^2}}. \tag{29}$$

We can conclude, using the union bound again, that for a $\log(n)$-balanced basis $\mathcal{B}$ the probability that there exist i_1, j_1, j_2 such that: $|s^{i_1,j_1} \cdot s^{i_1,j_2}| > \kappa$ is at most

$$p_{1,bad}(\kappa) \leq 2\sum_{i=1}^{m} \chi(i,i)e^{-\frac{\kappa^2}{2\xi^2(i,i)}\frac{n^2}{\log^4(n)}}. \tag{30}$$

Now let us find an upper bound on the expression $p_{2,bad}(\kappa) = \mathbb{P}[\exists_{i_1,i_2,j_1,j_2,i_1 \neq i_2} : |s^{i_1,j_1} \cdot s^{i_2,j_2}| > \kappa]$, where $i_1 \neq i_2$. We will assume that vectors of the basis $\mathcal{B}$ are $\log(n)$-balanced. Using the formula on $s^{i_1,j_1} \cdot s^{i_2,j_2}$ for $i_1 \neq i_2$, we get:

$$\mathbb{P}[|s^{i_1,j_1} \cdot s^{i_2,j_2}| > \kappa] = \mathbb{P}[|\sigma_{i_1,i_2}(1,1)x_1^{j_1}x_1^{j_2} + ... + \sigma_{i_1,i_2}(n,n)x_n^{j_1}x_n^{j_2} + 2\sum_{1 \leq n_1 < n_2 \leq n} d_{n_1}d_{n_2}x_{n_1}^{j_1}x_{n_2}^{j_2}\sigma_{i_1,i_2}(n_1,n_2)| > \kappa]. \tag{31}$$

Assume first that $P_i s$ are chosen deterministically. Note that by $\log(n)$-balanceness, we have:

$$|\sum_{n_1=1}^{n} \sigma_{i_1,i_2}(n_1,n_1)x_1^{j_1}x_1^{j_2}| \leq \frac{\log^2(n)}{n}\lambda(i_1,i_2). \tag{32}$$

Thus, by the triangle inequality, we have:

$$\mathbb{P}[|s^{i_1,j_1} \cdot s^{i_2,j_2}| > \kappa] \leq \mathbb{P}[|2\sum_{1 \leq n_1 < n_2 \leq n} d_{n_1}d_{n_2}x_{n_1}^{j_1}x_{n_2}^{j_2}\sigma_{i_1,i_2}(n_1,n_2)| \geq \kappa - \frac{\log^2(n)}{n}\lambda(i_1,i_2)]. \tag{33}$$

Using the same analysis as before, we then obtain the following bound on $p_{bad}(\kappa, \theta)$:

$$p_{2,bad}(\kappa) \leq 2\sum_{1 \leq i_1 < i_2 \leq m} \chi(i_1,i_2)e^{-\frac{(\kappa - \frac{\log^2(n)}{n}\lambda(i_1,i_2))^2}{2\xi^2(i,i)}\frac{n^2}{\log^4(n)}}. \tag{34}$$

We can conclude that in the setting where $P_i s$ are chosen deterministically, under our assumptions on $\lambda(i_1, i_2)$, for $\kappa > 0$ that does not depend on n and n large enough the following is true. The probability that there exist two different vector s^{i_1,j_1}, s^{i_2,j_2} such that $|s^{i_1,j_1} \cdot s^{i_2,j_2}| > \kappa$ satisfies:

$$p_{bad}(\kappa) \leq 2 \sum_{1 \leq i_1 \leq i_2 \leq m} \chi(i_1, i_2) e^{-\frac{(\kappa - \frac{\log^2(n)}{n}\lambda(i_1,i_2))^2}{2\xi^2(i,i)} \frac{n^2}{\log^4(n)}}. \tag{35}$$

Now let us assume that $P_i s$ are chosen probabilistically. In that setting we also assume that columns of different indices are chosen independently (this is the case for instance for the FastFood Transform). Let us now denote:

$$Y_j = \sigma_{i_1,i_2}(j,j) x_j^{j_1} x_j^{j^2} \tag{36}$$

for $j = 1, ..., n$. Denote $Y = \sum_{i=1}^{n} Y_1 + ... + Y_n$. Note that the condition on $\tilde{\lambda}(i_1, i_2)$ from the statement of the theorem implies that $\mathbb{E}[Y] = o_n(1)$. From the condition regarding independence of columns of different indices we deduce that $Y_i s$ are independent. Therefore we can apply Azuma's inequality and obtain the following bound on the expression: $\mathbb{P}[|Y - \mathbb{E}[Y]| > a]$:

$$\mathbb{P}[|Y - \mathbb{E}[Y]| > a] \leq 2e^{-\frac{a^2}{8\frac{\log^4(n)}{n^2}\sum_{j=1}^{n}(\sigma_{i_1,i_2}^{max}(j,j))^2}}. \tag{37}$$

If we now take $a = \frac{1}{\log(n)}$ and under $\log(n)$-balanceness assumption, we obtain:

$$\mathbb{P}[|Y - \mathbb{E}[Y]| > a] \leq 2e^{-\frac{n^2}{8\log^6(n)\sum_{j=1}^{n}(\sigma_{i_1,i_2}^{max}(j,j))^2}}. \tag{38}$$

Assume now that $|Y - \mathbb{E}[Y]| \leq \frac{1}{\log(n)}$. This happens with probability at least $1 - p_{wrong}$ with respect to the random choices of $P_i s$, where $p_{wrong} = 2e^{-\frac{n^2}{8\log^6(n)\sum_{j=1}^{n}(\sigma_{i_1,i_2}^{max}(j,j))^2}}$. But then random variable $|Y|$ is of the order $o_n(1)$.

Note that we have:

$$\mathbb{P}[|s^{i_1,j_1} \cdot s^{i_2,j_2}| > \kappa] = \mathbb{P}[|Y + 2\sum_{1 \leq n_1 < n_2 \leq n} d_{n_1} d_{n_2} x_{n_1}^{j_1} x_{n_2}^{j_2} \sigma_{i_1,i_2}(n_1, n_2)| > \kappa]. \tag{39}$$

Thus, using our bound on Y for a fixed κ and n large enough we can repeat previous analysis and conclude that in the probabilistic setting of $P_i s$ the following is true:

$$p_{bad}(\kappa) \leq 2 \sum_{1 \leq i_1 \leq i_2 \leq m} \chi(i_1, i_2) e^{-\frac{(\frac{\kappa}{2})^2}{2\xi^2(i,i)} \frac{n^2}{\log^4(n)}}. \tag{40}$$

Thus we can conclude that in both the deterministic and probabilistic setting for $P_i s$ we get:

$$p_{bad}(\kappa) \leq 2 \sum_{1 \leq i_1 \leq i_2 \leq m} \chi(i_1, i_2) e^{-\frac{\kappa^2}{8\xi^2(i,i)} \frac{n^2}{\log^4(n)}}. \tag{41}$$

Now we will show that the squared lengths of vectors $s^{i,j}$ are well concentrated around their means and that these means are equal to 1. Let us remind that we have:

$$s_l^{i,j} = d_1 p_{l,1}^i x_1^j + ... + d_n p_{l,n}^i x_n^j. \tag{42}$$

Thus we get:

$$\|s^{i,j}\|_2^2 = \sum_{1 \leq n_1 < n_2 \leq n} d_{n_1} d_{n_2} x_{n_1}^{j_1} x_{n_2}^{j_2} 2\sigma_{i,i}(n_1, n_2) + \sum_{n_1=1}^{n} (\sigma_{i,i}(n_1, n_1))^2 (x_{n_1}^j)^2 = \sum_{1 \leq n_1 < n_2 \leq n} d_{n_1} d_{n_2} x_{n_1}^{j_1} x_{n_2}^{j_2} 2\sigma_{i,i}(n_1, n_2) + 1, \tag{43}$$

where the last inequality comes from the fact that each column of each P_i has l_2-norm equal to 1.

Since obviously $\mathbb{E}[d_{n_1} d_{n_2} x_{n_1}^{j_1} x_{n_2}^{j_2} 2\sigma_{i,i}(n_1, n_2)] = 0$, then indeed $\mathbb{E}[\|s^{i,j}\|_2^2] = 1$. Let us find the upper bound on the following probability: $\mathbb{P}[|\|s^{i,j}\|_2^2 - 1| > \frac{1}{\log(n)}]$. We have:

$$\mathbb{P}[|\|s^{i,j}\|_2^2 - 1| > \frac{1}{\log(n)}] = \mathbb{P}[|d_{n_1} d_{n_2} x_{n_1}^{j_1} x_{n_2}^{j_2} 2\sigma_{i,i}(n_1, n_2)| > \frac{1}{\log(n)}]. \tag{44}$$

We can again apply Azuma's inequality and the union bound as we did before and obtain:

$$\mathbb{P}[\exists_{i,j} : |\|s^{i,j}\|_2^2 - 1| > \frac{1}{\log(n)}] \leq p_s, \tag{45}$$

where $p_s = 4\sum_{i=1}^{m} \chi(i,i) e^{-\frac{1}{2\xi^2(i,i)\log^2(n)} \frac{n^2}{\log^4(n)}}$.

We will assume now that all $s^{i,j}$ satisfy: $|\|s^{i,j}\|_2^2 - 1| \leq \frac{1}{\log(n)}$, in particular:

$$\sqrt{1 - \frac{1}{\log(n)}} \leq \|s^{i,j}\|_2 \leq \sqrt{1 + \frac{1}{\log(n)}}. \tag{46}$$

Let us assume right now that the above inequality holds. Let $\{w^{i,j}\}$ be a set of vectors obtained from $\{s^{i,j}\}$ by the Gram-Schmidt process. Without loss of generality we can assume that $\|w^{i,j}\|_2 = \|s^{i,j}\|_2$. Note that the size of the set $\{s^{i,j}\}$ is in fact not $2m$, but $2r$ and in all practical application $r \ll m$. Assume now that $|s^{i_1,j_1} \cdot s^{i_2,j_2}| \leq \kappa$ for any two different vectors s^{i_1,j_1}, s^{i_2,j_2} and some fixed $\kappa > 0$. Now, one can easily note that directly from the description of the Gram-Schmidt process that it leads to the set of vectors $\{w^{i,j}\}$ such that $\|s^{i,j} - w^{i,j}\|_2 \leq \kappa\Gamma(2r)$, where Γ is some constant that depends just on the size of the set $\{s^{i,j}\}$. Thus if we want ρ-orthogonality with $\rho = \frac{\epsilon}{\|g^{\mathcal{H}}\|_2}$, where $g^{\mathcal{H}}$ stands for the random projection of a vector g onto $2r$-dimensional linear space spanned by vectors from $\{s^{i,j}\}$, then we want to have:

$$\frac{\epsilon}{\|g^{\mathcal{H}}\|_2} = \kappa\Gamma(2r). \tag{47}$$

Thus we need to take:

$$\kappa = \frac{\epsilon}{\Gamma(2r)\|g^{\mathcal{H}}\|_2}. \tag{48}$$

Note that $g^{\mathcal{H}}$ is a $2r$-dimensional gaussian vector. Now let us take some $T > 0$. By the union bound the probability that $g^{\mathcal{H}}$ has l_2 norm greater than $\sqrt{2r} \cdot \sqrt{T}$ is at most: $2r\mathbb{P}[|\hat{g}|^2 > T]$, where $\hat{g}$ stands for a gaussian random variable taken from $\mathcal{N}(0,1)$. Now we use the following inequality for a tail of the gaussian random variable:

$$\mathbb{P}[|\hat{g}| > x] \leq 2\frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}}. \tag{49}$$

Thus we can conclude that the probability that $g^{\mathcal{H}}$ has l_2 norm larger than $\sqrt{2r} \cdot \sqrt{T}$ is at most $p_{gauss}(T) \leq \frac{4r}{\sqrt{2\pi T}}$. In such a case we need to take κ of the form:

$$\kappa = \frac{\epsilon}{\Gamma(2r)\sqrt{2r}\sqrt{T}}. \tag{50}$$

We are ready to finish the proof of Lemma 7.1. Take $\kappa = \frac{\epsilon}{\Gamma(2r)\sqrt{2r}\sqrt{T}}$. Let us first take the setting where $P_i s$ are chosen deterministically. Take an event $\mathcal{E}_{bad}$ which is the sum of the events which probabilisites are upper-bounded by $p_{gauss}(T)$, $1 - p_{balanced}$, $p_{bad}(\kappa)$ and p_s. By the union bound, the probability of that event is at most $p_{gauss} + (1 - p_{balanced}) + p_{bad}(\kappa) + p_s$ which is upper-bounded by $p_{gen} + p_{struct}$ for n large enough. Note that if $\mathcal{E}_{bad}$ does not hold then ρ-orthogonality is satisfied. Now let us take the probabilistic setting for choosing $P_i s$. We proceed similarly. The only difference is that right now we need to assume that the event upper-bounded by p_{wrong} does not hold (this one depends only on the random choices for setting up $P_i s$). Thus again we get the statement of the lemma. That completes the proof of Lemma 7.1. □

As mentioned above, the proof of Lemma 7.1 completes the proof of the theorem. □

Now we prove Theorem 4.2.

Proof. Fix some $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$. Assume that a matrix $\mathbf{A}$ is used to compute the approximation of the kernel $k(\mathbf{x}, \mathbf{z})$. Matrix $\mathbf{A}$ is either a truly random Gaussian matrix as it is the case in the unstructured computation or a structured matrix produced according to the $\mathcal{P}$-model. We assume that $\mathbf{A}$ has k rows and consists of $\frac{k}{m}$ blocks stacked vertically. If $\mathbf{A}$ is produced via the $\mathcal{P}$-model then each block is a structured matrix G^i_{struct}. The approximation of the kernel $\tilde{k}_{\mathcal{P}}(\mathbf{x}, \mathbf{z})$ is of the form: $\tilde{k}_{\mathbf{A}}(\mathbf{x}, \mathbf{z}) = \frac{1}{k}\sum_{i=1}^{\frac{k}{m}}\sum_{j=1}^{m}[\phi(a^{i,j} \cdot \mathbf{x}, a^{i,j} \cdot \mathbf{y})]$, where $a^{i,j}$ stands for the j^{th} row of the i^{th} block and $\phi : \mathbb{R}^2 \to \mathbb{R}$ is either of the form $\phi(a,b) = f(a)f(b)$, where f is a ReLU/sign function or $\phi(a,b) = \cos(a)\cos(b) + \sin(a)\sin(b)$. The latter formula for ϕ is valid if a kernel under consideration is Gaussian. Let use denote the random variable: $\phi(a^{i,j} \cdot \mathbf{x}, a^{i,j} \cdot \mathbf{y})$ as $X_{i,j}$. Then we have:

$$\tilde{k}_{\mathbf{A}}(\mathbf{x}, \mathbf{z}) = \frac{1}{k}\sum_{i=1}^{\frac{k}{m}}\sum_{j=1}^{m} X_{i,j}. \tag{51}$$

Thus we have:

$$\begin{aligned} Var(\tilde{k}_{\mathbf{A}}(\mathbf{x}, \mathbf{z})) = Var(\frac{1}{k}\sum_{i=1}^{\frac{k}{m}}\sum_{j=1}^{m} X_{i,j}) = \\ \frac{1}{k^2}Var(\sum_{i=1}^{\frac{k}{m}}\sum_{j=1}^{m} X_{i,j}) = \frac{1}{k^2}[\sum_{i=1}^{\frac{k}{m}}\sum_{j=1}^{m} Var(X_{i,j}) + \\ \sum_{i,j_1 \neq j_2} Cov(X_{i,j_1}, X_{i,j_2})]. \end{aligned} \tag{52}$$

The last inequality in Eqn.52 is implied by the fact that different blocks of the structured matrix are computed independently and thus covariance related to rows from different blocks is 0.

Therefore we obtain:

$$\begin{aligned} Var(\tilde{k}_{\mathbf{A}}(\mathbf{x}, \mathbf{z})) = \frac{1}{k^2}\sum_{i=1}^{\frac{k}{m}}\sum_{j=1}^{m} Var(X_{i,j}) + \\ \frac{1}{k^2}\sum_{i,j_1 \neq j_2}(\mathbb{E}[X_{i,j_1}, X_{i,j_2}] - \mathbb{E}[X_{i,j_1}]\mathbb{E}[X_{i,j_2}]). \end{aligned} \tag{53}$$

Now note that the first expression on the RHS above is the same for both the structured and unstructured setting. This is the case since one can note that $X_{i,j}$ has the same distribution in the unstructured and structured setting. For the same reason the expression $\mathbb{E}[X_{i,j_1}]\mathbb{E}[X_{i,j_2}]$ is the same for the structured and unstructured setting. Thus if $\mathbf{G}$ stands for the fully unstructured model and we denote $\tilde{k}_{\mathbf{A}}(\mathbf{x}, \mathbf{z}) = \tilde{k}_{\mathcal{P}}(\mathbf{x}, \mathbf{z})$ if A is constructed according to the $\mathcal{P}$-model, then we get:

$$\begin{aligned} |Var(\tilde{k}_{\mathbf{G}}(\mathbf{x}, \mathbf{z})) - Var(\tilde{k}_{\mathcal{P}}(\mathbf{x}, \mathbf{z}))| \leq \\ \frac{1}{k^2}\sum_{i,j_1 \neq j_2}|\mathbb{E}[X^{\mathcal{P}}_{i,j_1}X^{\mathcal{P}}_{i,j_2}] - \mathbb{E}[X^{\mathbf{G}}_{i,j_1}X^{\mathbf{G}}_{i,j_2}]|, \end{aligned} \tag{54}$$

where $X_{i,j}^{\mathcal{P}}$ stands for the version of $X_{i,j}$ if $\mathbf{A}$ was costructed via the $\mathcal{P}$-model and $X_{i,j}^{\mathbf{G}}$ stands for the fully unstructured one.

Therfore we have:

$$|Var(\tilde{k}_{\mathbf{G}}(\mathbf{x},\mathbf{z})) - Var(\tilde{k}_{\mathcal{P}}(\mathbf{x},\mathbf{z}))| \leq \frac{1}{k^2} \cdot \frac{k}{m} \sum_{j_1 \neq j_2} |\mathbb{E}[X_{1,j_1}^{\mathcal{P}} X_{1,j_2}^{\mathcal{P}}] - \mathbb{E}[X_{1,j_1}^{\mathbf{G}} X_{1,j_2}^{\mathbf{G}}]|, \tag{55}$$

where the latter inequality is implied by the fact that different blocks are constructed independently.

Therefore we get:

$$|Var(\tilde{k}_{\mathbf{G}}(\mathbf{x},\mathbf{z})) - Var(\tilde{k}_{\mathcal{P}}(\mathbf{x},\mathbf{z}))| \leq \frac{1}{k^2} \cdot \frac{k}{m} \binom{m}{2} \beta, \tag{56}$$

where β is an upper bound as in Theorem 7.1 for $d = 2$. Now we can proceed in the same way as in the proof of Theorem 4.1 and the proof is completed. □

Now we prove Theorem 4.3.

Proof. The fact that $\mu[\mathcal{P}] \leq \kappa$ comes directly from the definition of the coherence number and the sparse setting of semi-gaussian matrices. To see that, note that any given column col of any matrix $\mathbf{P}_i$ in the related $\mathcal{P}$-model has a nonzero dot-product with at most κ^2 other columns of any matrix $\mathbf{P}_j$. This in turn is implied by the fact that different columns are obtained by applying skew-circulant shifts blockwise, thus the number of columns from $\mathbf{P}_j$ that have nonzero dot product with col is at most the product of the number of nonzero dimensions of col and $\mathbf{P}_j$. This is clearly upper bounded by κ^2. This leads to the upper bound on the coherence $\mu[\mathcal{P}]$.

The new formula for p_{wrong} is derived by a similar analysis to the one used to obtain the formula on p_{wrong} in the proof of Theorem 4.1. This time random variables under analysis are not independent though, but using the same trick as the one we used in the proof of Theorem 4.1 to decouple dependent random variables in the sum to be estimated and applying Azuma's inequality (we omit details since the analysis is exactly the same as in the aforementioned proof), we obtain the following: $\mathbb{P}[|\mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_1}| > c] \leq e^{-\Omega(rc^2)}$ for $i \neq j$ and any constant $c > 0$. Taking the union bound over all the pairs of columns and fixing $c = \frac{1}{\log^2(n)}$ and $r = 3\log^5(n)$, we can conclude that with probability at least $1 - o(\frac{1}{n})$ the absolute value of the expression $\lambda(i,j)$ from the proof of Theorem 4.1 is of the order $o(\frac{n}{\log^2(n)})$. That enables us to finish tha analysis in the same way as in the proof of Theorem 4.1 and derive similar conclusions.

The bound regarding the chromatic number is implied by the observation that each coherence graph in the corresponding $\mathcal{P}$-model has degree at most κ^2. That follows directly from the observation we used to prove the upper bound on $\mu[\mathcal{P}]$. But now we can use Lemma 3.1 and that completes the proof of Theorem 4.3. □

Below we present the proof of Theorem 4.4.

Proof. Fix two columns $\mathbf{P}_{i,n_1}$ and $\mathbf{P}_{j,n_2}$ and consider the expression $\mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_2}$. We have already mentioned in the previous proof the right approach to finding strong upper bound on $|\mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_2}|$. We first note that $\mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_2}$ can be written as a sum $w_1 + ... + w_{nr}$, where $w_i s$ are not necessarily independent but can be partitioned into at most three sets such that wariables in each of these sets are independent. This is true since G_{struct}^i is produced by skew-circulant shifts and the corresponding coherence graphs has verrtices of degree at most 2. Note also that each w_k satisfies: $|w_k| \leq \frac{1}{\alpha r}$. In each of the sum we get rid of these $w_i s$ that are equal to 0. Then, by applying Azuma's inequality independently on each of these subsets and taking union bound over these subsets, we conclude that for any $a > 0$:

$$|\mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_2} > a| \leq 3e^{-\frac{a^2 \alpha r}{O(1)}} \tag{57}$$

Now we can take the union bound over all pairs of columns and notice that for every columcn col in $\mathbf{P}_i$ and any $\mathbf{P}_j$ there exists at most κ columns in $\mathbf{P}_j$ that have nonzero dot product with col. We can then take $a = \frac{\tau}{\kappa}$ and the proof is completed. □

Let us now switch to dense semi-gaussian matrices. The following is true.

Theorem 7.2. *Consider the setting as in Theorem 4.1. Assume that entries of any fixed column of P_i are chosen independently at random. Assume also that for any $1 \leq i \leq j \leq m$ and any fixed column col of P_i each column of P_j is a downward shift of col by b entries (possibly with signs of dimensions swapped) and that $b = 0$ for $O(1)$ columns in P_j. Then for and $T > 0$ and n large enough the following holds:*

$$|\mathbb{E}[\tilde{k}_{\mathcal{P}}^d(\boldsymbol{x},\boldsymbol{z})] - \mathbb{E}[\tilde{k}_{\mathbf{G}}^d(\boldsymbol{x},\boldsymbol{z})]| \leq O(\Delta), \tag{58}$$

where $\Delta = p_{gen}(T) + p_{struct}(T) + d\epsilon + e^{-n^{\frac{1}{3}}}$ *and*

$$\epsilon = \frac{\log^3(n)}{n} \left(n^{\frac{2}{3}} + \max_{1 \leq i \leq j \leq m} |\sum_{1 \leq n_1 < n_2 \leq n} \mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_2}| \right).$$

As a corollary:

$$|Var(\tilde{k}_{\mathcal{P}}(\boldsymbol{x},\boldsymbol{z})) - Var(\tilde{k}_{\mathbf{G}}(\boldsymbol{x},\boldsymbol{z}))| = O(\frac{m-1}{2k}\Delta). \tag{59}$$

Proof. The proof of this result follows along the lines of the proof of Theorem 4.1 and Theorem 4.2. Take the formulas for $s^{i_1,j_1} \cdot s^{i_2,j_2}$ derived in the proof of Theorem 7.1. Note that we want to have: $|s^{i_1,j_1} \cdot s^{i_2,j_2}| \leq \frac{\epsilon}{\Gamma(2d)\|g^{\mathcal{H}}\|_2}$, where Γ is a constant that depends only on the degree d. Each $s^{i_1,j_1} \cdot s^{i_2,j_2}$ is a sum of random variables that can be decoupled into $O(1)$ subsums such that variables in each subsum are independent (here we use exactly the same trick as in the proof of Theorem 4.1). In each subsum we apply Azuma's inequality. Straightforward computations lead to the conclusion that if one sets up ϵ as in the statement of Theorem 7.2 then the probability that there exist different s^{i_1,j_1}, s^{i_2,j_2} such that $|s^{i_1,j_1} \cdot s^{i_2,j_2}| > \frac{\epsilon}{\Gamma(2d)\|g^{\mathcal{H}}\|_2}$ is of the order $e^{-n^{\frac{1}{3}}}$ for n large enough. That is the extra term in the formula for Δ that was not present in the staement of Theorem 4.1. The variance results follows immediately by exactly the same analysis as in the proof of Theorem 4.2. □

Note that introduced dense semi-gaussian matrices trivially satisfy conditions of Theorem 7.2 (look for the description of matrices $\mathbf{P}_i$ from Subsection: 3.2.4). The role of rank is similar as in the sparse setting, i.e. larger values of r lead to sharper concentration results. Theorem 7.2 can be applied to classes of matrices for which $|\sum_{1 \leq n_1 < n_2 \leq n} \mathbf{P}_{i,n_1}^T \mathbf{P}_{j,n_2}|$ is small and random dense semi-gaussian matrices satisfy this condition with high probability.