The Unreasonable Effectiveness of Structured Random Orthogonal Embeddings

Krzysztof Choromanski *
Google Brain Robotics
kchoro@google.com

Mark Rowland *
University of Cambridge
mr504@cam.ac.uk

Adrian Weller
University of Cambridge and Alan Turing Institute
aw665@cam.ac.uk

Abstract

We examine a class of embeddings based on structured random matrices with orthogonal rows which can be applied in many machine learning applications including dimensionality reduction and kernel approximation. For both the Johnson-Lindenstrauss transform and the angular kernel, we show that we can select matrices yielding guaranteed improved performance in accuracy and/or speed compared to earlier methods. We introduce matrices with complex entries which give significant further accuracy improvement. We provide geometric and Markov chain-based perspectives to help understand the benefits, and empirical results which suggest that the approach is helpful in a wider range of applications.

1 Introduction

Embedding methods play a central role in many machine learning applications by projecting feature vectors into a new space (often nonlinearly), allowing the original task to be solved more efficiently. The new space might have more or fewer dimensions depending on the goal. Applications include the Johnson-Lindenstrauss Transform for dimensionality reduction (JLT, Johnson and Lindenstrauss, 1984) and kernel methods with random feature maps (Rahimi and Recht, 2007). The embedding can be costly hence many fast methods have been developed, see §1.1 for background and related work.

We present a general class of random embeddings based on particular structured random matrices with orthogonal rows, which we call *random ortho-matrices* (ROMs); see §2. We show that ROMs may be used for the applications above, in each case demonstrating improvements over previous methods in statistical accuracy (measured by mean squared error, MSE), in computational efficiency (while providing similar accuracy), or both. We highlight the following contributions:

- In §3: The *Orthogonal Johnson-Lindenstrauss Transform* (OJLT) for dimensionality reduction. We prove this has strictly smaller MSE than the previous unstructured JLT mechanisms. Further, OJLT is as fast as the fastest previous JLT variants (which are structured).
- In §4: Estimators for the *angular kernel* (Sidorov et al., 2014) which guarantee better MSE. The *angular kernel* is important for many applications, including natural language processing (Sidorov et al., 2014), image analysis (Jégou et al., 2011), speaker representations (Schmidt et al., 2014) and tf-idf data sets (Sundaram et al., 2013).
- In §5: Two perspectives on the effectiveness of ROMs to help build intuitive understanding.

In §6 we provide empirical results which support our analysis, and show that ROMs are effective for a still broader set of applications. Full details and proofs of all results are in the Appendix.

*equal contribution

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

1.1 Background and related work

Our ROMs can have two forms (see §2 for details): (i) a $\mathbf{G}_{\mathrm{ort}}$ is a random Gaussian matrix conditioned on rows being orthogonal; or (ii) an $\mathbf{SD}$-*product* matrix is formed by multiplying some number k of $\mathbf{SD}$ blocks, each of which is highly structured, typically leading to fast computation of products. Here $\mathbf{S}$ is a particular structured matrix, and $\mathbf{D}$ is a random diagonal matrix; see §2 for full details. Our $\mathbf{SD}$ block generalizes an $\mathbf{HD}$ block, where $\mathbf{H}$ is a *Hadamard* matrix, which received previous attention. Earlier approaches to embeddings have explored using various structured matrices, including particular versions of one or other of our two forms, though in different contexts.

For dimensionality reduction, Ailon and Chazelle (2006) used a single $\mathbf{HD}$ block as a way to spread out the mass of a vector over all dimensions before applying a sparse Gaussian matrix. Choromanski and Sindhwani (2016) also used just one $\mathbf{HD}$ block as part of a larger structure. Bojarski et al. (2017) discussed using $k = 3$ $\mathbf{HD}$ blocks for locality-sensitive hashing methods but gave no concrete results for their application to dimensionality reduction or kernel approximation. All these works, and other earlier approaches (Hinrichs and Vybíral, 2011; Vybíral, 2011; Zhang and Cheng, 2013; Le et al., 2013; Choromanska et al., 2016), provided computational benefits by using structured matrices with less randomness than unstructured iid Gaussian matrices, but none demonstrated accuracy gains.

Yu et al. (2016) were the first to show that $\mathbf{G}_{\mathrm{ort}}$-type matrices can yield improved accuracy, but their theoretical result applies only asymptotically for many dimensions, only for the Gaussian kernel and for just one specific orthogonal transformation, which is one instance of the larger class we consider. Their theoretical result does not yield computational benefits. Yu et al. (2016) did explore using a number k of $\mathbf{HD}$ blocks empirically, observing good computational and statistical performance for $k = 3$, but without any theoretical accuracy guarantees. It was left as an open question why matrices formed by a small number of $\mathbf{HD}$ blocks can outperform non-discrete transforms.

In contrast, we are able to prove that ROMs yield improved MSE in several settings and for many of them for any number of dimensions. In addition, $\mathbf{SD}$-product matrices can deliver computational speed benefits. We provide initial analysis to understand why $k = 3$ can outperform the state-of-the-art, why odd k yields better results than even k, and why higher values of k deliver decreasing additional benefits (see §3 and §5).

2 The family of Random Ortho-Matrices (ROMs)

Random ortho-matrices (ROMs) are taken from two main classes of distributions defined below that require the rows of sampled matrices to be orthogonal. A central theme of the paper is that this orthogonal structure can yield improved statistical performance. We shall use bold uppercase (e.g. $\mathbf{M}$) to denote matrices and bold lowercase (e.g. $\mathbf{x}$) for vectors.

Gaussian orthogonal matrices. Let $\mathbf{G}$ be a random matrix taking values in $\mathbb{R}^{m\times n}$ with iid $\mathcal{N}(0,1)$ elements, which we refer to as an *unstructured* Gaussian matrix. The first ROM distribution we consider yields the random matrix $\mathbf{G}_{\mathrm{ort}}$, which is defined as a random $\mathbb{R}^{n\times n}$ matrix given by first taking the rows of the matrix to be a uniformly random orthonormal basis, and then independently scaling each row, so that the rows marginally have multivariate Gaussian $\mathcal{N}(0, I)$ distributions. The random variable $\mathbf{G}_{\mathrm{ort}}$ can then be extended to non-square matrices by either stacking independent copies of the $\mathbb{R}^{n\times n}$ random matrices, and deleting superfluous rows if necessary. The orthogonality of the rows of this matrix has been observed to yield improved statistical properties for randomized algorithms built from the matrix in a variety of applications.

SD-product matrices. Our second class of distributions is motivated by the desire to obtain similar statistical benefits of orthogonality to $\mathbf{G}_{\mathrm{ort}}$, whilst gaining computational efficiency by employing more structured matrices. We call this second class $\mathbf{SD}$-*product* matrices. These take the more structured form $\prod_{i=1}^{k}\mathbf{SD}_i$, where $\mathbf{S} = \{s_{i,j}\} \in \mathbb{R}^{n\times n}$ has orthogonal rows, $|s_{i,j}| = \frac{1}{\sqrt{n}}$ $\forall i, j \in \{1,\ldots,n\}$; and the $(\mathbf{D}_i)_{i=1}^{k}$ are independent diagonal matrices described below. By $\prod_{i=1}^{k}\mathbf{SD}_i$, we mean the matrix product $(\mathbf{SD}_k)\ldots(\mathbf{SD}_1)$. This class includes as particular cases several recently introduced random matrices (e.g. Andoni et al., 2015; Yu et al., 2016), where good *empirical* performance was observed. We go further to establish strong theoretical guarantees, see §3 and §4.

A prominent example of an $\mathbf{S}$ matrix is the normalized *Hadamard* matrix $\mathbf{H}$, defined recursively by $\mathbf{H}_1 = (1)$, and then for $i > 1$, $\mathbf{H}_i = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{H}_{i-1} & \mathbf{H}_{i-1} \\ \mathbf{H}_{i-1} & -\mathbf{H}_{i-1} \end{pmatrix}$. Importantly, matrix-vector products with $\mathbf{H}$ are computable in $O(n \log n)$ time via the fast Walsh-Hadamard transform, yielding large computational savings. In addition, $\mathbf{H}$ matrices enable a significant space advantage: since the fast Walsh-Hadamard transform can be computed without explicitly storing $\mathbf{H}$, only $O(n)$ space is required to store the diagonal elements of $(\mathbf{D}_i)_{i=1}^k$. Note that these $\mathbf{H}_n$ matrices are defined only for n a power of 2, but if needed, one can always adjust data by padding with 0s to enable the use of 'the next larger' $\mathbf{H}$, doubling the number of dimensions in the worst case.

Matrices $\mathbf{H}$ are representatives of a much larger family in $\mathbf{S}$ which also attains computational savings. These are L_2-normalized versions of Kronecker-product matrices of the form $\mathbf{A}_1 \otimes ... \otimes \mathbf{A}_l \in \mathbb{R}^{n \times n}$ for $l \in \mathbb{N}$, where $\otimes$ stands for a Kronecker product and blocks $\mathbf{A}_i \in \mathbb{R}^{d \times d}$ have entries of the same magnitude and pairwise orthogonal rows each. For these matrices, matrix-vector products are computable in $O(n(2d-1)\log_d(n))$ time (Zhang et al., 2015).

$\mathbf{S}$ includes also the *Walsh matrices* $\mathbf{W} = \{w_{i,j}\} \in \mathbb{R}^{n \times n}$, where $w_{i,j} = \frac{1}{\sqrt{n}}(-1)^{i_{N-1}j_0 + ... + i_0 j_{N-1}}$ and $i_{N-1}...i_0$, $j_{N-1}...j_0$ are binary representations of i and j respectively.

For diagonal $(\mathbf{D}_i)_{i=1}^k$, we mainly consider Rademacher entries leading to the following matrices.

Definition 2.1. *The $\mathbf{S}$-Rademacher random matrix with $k \in \mathbb{N}$ blocks is below, where $(\mathbf{D}_i^{(\mathcal{R})})_{i=1}^k$ are diagonal with iid Rademacher random variables [i.e. $\mathrm{Unif}(\{\pm 1\})$] on the diagonals:*

$$\mathbf{M}_{\mathbf{S}\mathcal{R}}^{(k)} = \prod_{i=1}^{k} \mathbf{S}\mathbf{D}_i^{(\mathcal{R})}. \tag{1}$$

Having established the two classes of ROMs, we next apply them to dimensionality reduction.

3 The Orthogonal Johnson-Lindenstrauss Transform (OJLT)

Let $\mathcal{X} \subset \mathbb{R}^n$ be a dataset of n-dimensional real vectors. The goal of dimensionality reduction via random projections is to transform linearly each $\mathbf{x} \in \mathcal{X}$ by a random mapping $\mathbf{x} \overset{F}{\mapsto} \mathbf{x}'$, where: $F : \mathbb{R}^n \to \mathbb{R}^m$ for $m < n$, such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ the following holds: $(\mathbf{x}')^\top \mathbf{y}' \approx \mathbf{x}^\top \mathbf{y}$. If we furthermore have $\mathbb{E}[(\mathbf{x}')^\top \mathbf{y}'] = \mathbf{x}^\top \mathbf{y}$ then the dot-product estimator is *unbiased*. In particular, this dimensionality reduction mechanism should in expectation preserve information about vectors' norms, i.e. we should have: $\mathbb{E}[\|\mathbf{x}'\|_2^2] = \|\mathbf{x}\|_2^2$ for any $\mathbf{x} \in \mathcal{X}$.

The standard JLT mechanism uses the randomized linear map $F = \frac{1}{\sqrt{m}}\mathbf{G}$, where $\mathbf{G} \in \mathbb{R}^{m \times n}$ is as in §2, requiring mn multiplications to evaluate. Several fast variants (FJLTs) have been proposed by replacing $\mathbf{G}$ with random structured matrices, such as sparse or circulant Gaussian matrices (Ailon and Chazelle, 2006; Hinrichs and Vybíral, 2011; Vybíral, 2011; Zhang and Cheng, 2013). The fastest of these variants has $O(n \log n)$ time complexity, but at a cost of higher MSE for dot-products.

Our Orthogonal Johnson-Lindenstrauss Transform (OJLT) is obtained by replacing the unstructured random matrix $\mathbf{G}$ with a sub-sampled ROM from §2: either $\mathbf{G}_{\mathrm{ort}}$, or a sub-sampled version $\mathbf{M}_{\mathbf{S}\mathcal{R}}^{(k),\mathrm{sub}}$ of the $\mathbf{S}$-Rademacher ROM, given by sub-sampling rows from the left-most $\mathbf{S}$ matrix in the product. We sub-sample since $m < n$. We typically assume uniform sub-sampling *without* replacement. The resulting dot-product estimators for vectors $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ are given by:

$$\widehat{K}_m^{\mathrm{base}}(\mathbf{x}, \mathbf{y}) = \frac{1}{m}(\mathbf{G}\mathbf{x})^\top(\mathbf{G}\mathbf{y}) \quad \text{[unstructured iid baseline, previous state-of-the-art accuracy]},$$

$$\widehat{K}_m^{\mathrm{ort}}(\mathbf{x}, \mathbf{y}) = \frac{1}{m}(\mathbf{G}_{\mathrm{ort}}\mathbf{x})^\top(\mathbf{G}_{\mathrm{ort}}\mathbf{y}), \qquad \widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y}) = \frac{1}{m}\left(\mathbf{M}_{\mathbf{S}\mathcal{R}}^{(k),\mathrm{sub}}\mathbf{x}\right)^\top \left(\mathbf{M}_{\mathbf{S}\mathcal{R}}^{(k),\mathrm{sub}}\mathbf{y}\right). \tag{2}$$

We contribute the following closed-form expressions, which exactly quantify the mean-squared error (MSE) for these three estimators. Precisely, the MSE of an estimator $\widehat{K}(\mathbf{x}, \mathbf{y})$ of the inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ for $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ is defined to be $\mathrm{MSE}(\widehat{K}(\mathbf{x}, \mathbf{y})) = \mathbb{E}\left[(\widehat{K}(\mathbf{x}, \mathbf{y}) - \langle \mathbf{x}, \mathbf{y} \rangle^2)\right]$. See the Appendix for detailed proofs of these results and all others in this paper.

Lemma 3.1. *The MSE of the unstructured JLT dot-product estimator $\widehat{K}_m^{\mathrm{base}}$ of $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ using m-dimensional random feature maps is unbiased, with* $\mathrm{MSE}(\widehat{K}_m^{\mathrm{base}}(\mathbf{x}, \mathbf{y})) = \frac{1}{m}((\mathbf{x}^\top \mathbf{y})^2 + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2)$.

Theorem 3.2. *The estimator $\widehat{K}_m^{\mathrm{ort}}$ is unbiased and satisfies, for $n \geq 4$:*

$$\begin{aligned}
&\mathrm{MSE}(\widehat{K}_m^{\mathrm{ort}}(\mathbf{x}, \mathbf{y})) \\
=&\mathrm{MSE}(\widehat{K}_m^{\mathrm{base}}(\mathbf{x}, \mathbf{y})) + \\
&\frac{m}{m-1}\left[\frac{\|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 n^2}{4I(n-3)I(n-4)}\left(\left(\frac{1}{n} - \frac{1}{n+2}\right)(I(n-3) - I(n-1))I(n-4)\left[\cos^2(\theta) + \frac{1}{2}\right] + \right.\right. \\
&\left.\left. I(n-1)\left(I(n-4) - I(n-2)\right)\left(\frac{1}{n-2} - \frac{1}{n}\right)\left[\cos^2(\theta) - \frac{1}{2}\right]\right) - \langle \mathbf{x}, \mathbf{y} \rangle^2\right],
\end{aligned} \tag{3}$$

where $I(n) = \int_0^\pi \sin^n(x) dx = \frac{\sqrt{\pi}\Gamma((n+1)/2)}{\Gamma(n/2+1)}$.

Theorem 3.3 (Key result)**.** *The OJLT estimator $\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y})$ with k blocks, using m-dimensional random feature maps and uniform sub-sampling policy without replacement, is unbiased with*

$$\begin{aligned}
\mathrm{MSE}(\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y})) = \frac{1}{m}\left(\frac{n-m}{n-1}\right)\Bigg(&((\mathbf{x}^\top \mathbf{y})^2 + \|\mathbf{x}\|^2 \|\mathbf{y}\|^2) + \\
&\sum_{r=1}^{k-1} \frac{(-1)^r 2^r}{n^r}(2(\mathbf{x}^\top \mathbf{y})^2 + \|\mathbf{x}\|^2 \|\mathbf{y}\|^2) + \frac{(-1)^k 2^k}{n^{k-1}} \sum_{i=1}^{n} x_i^2 y_i^2\Bigg).
\end{aligned} \tag{4}$$

Proof (Sketch). For $k = 1$, the random projection matrix is given by sub-sampling rows from $\mathbf{SD}_1$, and the computation can be carried out directly. For $k \geq 1$, the proof proceeds by induction. The random projection matrix in the general case is given by sub-sampling rows of the matrix $\mathbf{SD}_k \cdots \mathbf{SD}_1$. By writing the MSE as an expectation and using the law of conditional expectations conditioning on the value of the first $k-1$ random matrices $\mathbf{D}_{k-1}, \ldots, \mathbf{D}_1$, the statement of the theorem for 1 $\mathbf{SD}$ block and for $k-1$ $\mathbf{SD}$ blocks can be neatly combined to yield the result. □

To our knowledge, it has not previously been possible to provide theoretical guarantees that $\mathbf{SD}$-product matrices outperform iid matrices. Combining Lemma 3.1 with Theorem 3.3 yields the following important result.

Corollary 3.4 (Theoretical guarantee of improved performance)**.** *Estimators $\widehat{K}_m^{(k)}$ (subsampling without replacement) yield guaranteed lower MSE than $\widehat{K}_m^{\mathrm{base}}$.*

It is not yet clear when $\widehat{K}_m^{\mathrm{ort}}$ is better or worse than $\widehat{K}_m^{(k)}$; we explore this empirically in §6. Theorem 3.3 shows that there are diminishing MSE benefits to using a large number k of $\mathbf{SD}$ blocks. Interestingly, odd k is better than even: it is easy to observe that $\mathrm{MSE}(\widehat{K}_m^{(2k-1)}(\mathbf{x}, \mathbf{y})) < \mathrm{MSE}(\widehat{K}_m^{(2k)}(\mathbf{x}, \mathbf{y})) > \mathrm{MSE}(\widehat{K}_m^{(2k+1)}(\mathbf{x}, \mathbf{y}))$. These observations, and those in §5, help to understand why empirically $k = 3$ was previously observed to work well (Yu et al., 2016).

If we take $\mathbf{S}$ to be a normalized Hadamard matrix $\mathbf{H}$, then even though we are using sub-sampling, and hence the full computational benefits of the Walsh-Hadamard transform are not available, still $\widehat{K}_m^{(k)}$ achieves improved MSE compared to the base method with *less* computational effort, as follows.

Lemma 3.5. *There exists an algorithm (see Appendix for details) which computes an embedding for a given datapoint $\mathbf{x}$ using $\widehat{K}_m^{(k)}$ with $\mathbf{S}$ set to $\mathbf{H}$ and uniform sub-sampling policy in expected time* $\min\{O((k-1)n\log(n) + nm - \frac{(m-1)m}{2}, kn\log(n)\}$.

Note that for $m = \omega(k\log(n))$ or if $k = 1$, the time complexity is smaller than the brute force $\Theta(nm)$. The algorithm uses a simple observation that one can reuse calculations conducted for the upper half of the Hadamard matrix while performing computations involving rows from its other half, instead of running these calculations from scratch (details in the Appendix).

An alternative to sampling without replacement is deterministically to choose the first m rows. In our experiments in §6, these two approaches yield the same empirical performance, though we expect

that the deterministic method could perform poorly on adversarially chosen data. The first m rows approach can be realized in time $O(n\log(m)+(k-1)n\log(n))$ per datapoint.

Theorem 3.3 is a key result in this paper, demonstrating that $\mathbf{SD}$-product matrices yield both statistical and computational improvements compared to the base iid procedure, which is widely used in practice. We next show how to obtain further gains in accuracy.

3.1 Complex variants of the OJLT

We show that the MSE benefits of Theorem 3.3 may be markedly improved by using $\mathbf{SD}$-product matrices with complex entries $\mathbf{M}_{\mathbf{S}\mathcal{H}}^{(k)}$. Specifically, we consider the variant $\mathbf{S}$-*Hybrid* random matrix below, where $\mathbf{D}_k^{(\mathcal{U})}$ is a diagonal matrix with iid $\mathrm{Unif}(S^1)$ random variables on the diagonal, independent of $(\mathbf{D}_i^{(\mathcal{R})})_{i=1}^{k-1}$, and S^1 is the unit circle of $\mathbb{C}$. We use the real part of the Hermitian product between projections as a dot-product estimator; recalling the definitions of §2, we use:

$$\mathbf{M}_{\mathbf{S}\mathcal{H}}^{(k)}=\mathbf{S}\mathbf{D}_k^{(\mathcal{U})}\prod_{i=1}^{k-1}\mathbf{S}\mathbf{D}_i^{(\mathcal{R})}\,,\qquad \widehat{K}_m^{\mathcal{H},(k)}(\mathbf{x},\mathbf{y})=\frac{1}{m}\,\mathrm{Re}\left[\left(\overline{\mathbf{M}_{\mathbf{S}\mathcal{H}}^{(k),\mathrm{sub}}\mathbf{x}}\right)^{\top}\left(\mathbf{M}_{\mathbf{S}\mathcal{H}}^{(k),\mathrm{sub}}\mathbf{y}\right)\right]. \quad (5)$$

Remarkably, this complex variant yields exactly half the MSE of the OJLT estimator.

Theorem 3.6. *The estimator $\widehat{K}_m^{\mathcal{H},(k)}(\mathbf{x},\mathbf{y})$, applying uniform sub-sampling without replacement, is unbiased and satisfies:* $\mathrm{MSE}(\widehat{K}_m^{\mathcal{H},(k)}(\mathbf{x},\mathbf{y}))=\frac{1}{2}\mathrm{MSE}(\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y}))$.

This large factor of 2 improvement could instead be obtained by doubling m for $\widehat{K}_m^{(k)}$. However, this would require doubling the number of parameters for the transform, whereas the $\mathbf{S}$-Hybrid estimator requires additional storage only for the complex parameters in the matrix $\mathbf{D}_k^{(\mathcal{U})}$. Strikingly, it is straightforward to extend the proof of Theorem 3.6 (see Appendix) to show that rather than taking the complex random variables in $\mathbf{M}_{\mathbf{S}\mathcal{H}}^{(k),\mathrm{sub}}$ to be $\mathrm{Unif}(S^1)$, it is possible to take them to be $\mathrm{Unif}(\{1,-1,i,-i\})$ and still obtain exactly the same benefit in MSE.

Theorem 3.7. *For the estimator $\widehat{K}_m^{\mathcal{H},(k)}$ defined in Equation* (5)*: replacing the random matrix $\mathbf{D}_k^{(\mathcal{U})}$ (which has iid $\mathrm{Unif}(S^1)$ elements on the diagonal) with instead a random diagonal matrix having iid $\mathrm{Unif}(\{1,-1,i,-i\})$ elements on the diagonal, does not affect the MSE of the estimator.*

It is natural to wonder if using an $\mathbf{SD}$-product matrix with more complex random variables (for all $\mathbf{SD}$ blocks) would improve performance still further. However, interestingly, this appears not to be the case; details are provided in the Appendix §8.7.

3.2 Sub-sampling with replacement

Our results above focus on $\mathbf{SD}$-product matrices where rows have been sub-sampled without replacement. Sometimes (e.g. for parallelization) it can be convenient instead to sub-sample *with* replacement. As might be expected, this leads to worse MSE, which we can quantify precisely.

Theorem 3.8. *For each of the estimators $\widehat{K}_m^{(k)}$ and $\widehat{K}_m^{\mathcal{H},(k)}$, if uniform sub-sampling* with *(rather than without) replacement is used then the MSE is worsened by a multiplicative constant of $\frac{n-1}{n-m}$.*

4 Kernel methods with ROMs

ROMs can also be used to construct high-quality random feature maps for non-linear kernel approximation. We analyze here the *angular kernel*, an important example of a *Pointwise Nonlinear Gaussian kernel* (PNG), discussed in more detail at the end of this section.

Definition 4.1. *The angular kernel K^{ang} is defined on $\mathbb{R}^n$ by $K^{\mathrm{ang}}(\mathbf{x},\mathbf{y})=1-\frac{2\theta_{\mathbf{x},\mathbf{y}}}{\pi}$, where $\theta_{\mathbf{x},\mathbf{y}}$ is the angle between $\mathbf{x}$ and $\mathbf{y}$.*

To employ random feature style approximations to this kernel, we first observe it may be rewritten as

$$K^{\mathrm{ang}}(\mathbf{x},\mathbf{y}) = \mathbb{E}\left[\mathrm{sign}(\mathbf{G}\mathbf{x})\mathrm{sign}(\mathbf{G}\mathbf{y})\right] ,$$

where $\mathbf{G} \in \mathbb{R}^{1\times n}$ is an unstructured isotropic Gaussian vector. This motivates approximations of the form:

$$\widehat{K}^{\mathrm{ang}}m(\mathbf{x},\mathbf{y}) = \frac{1}{m}\mathrm{sign}(\mathbf{M}\mathbf{x})^{\top}\mathrm{sign}(\mathbf{M}\mathbf{y}), \tag{6}$$

where $\mathbf{M} \in \mathbb{R}^{m\times n}$ is a random matrix, and the sign function is applied coordinate-wise. Such kernel estimation procedures are heavily used in practice (Rahimi and Recht, 2007), as they allow fast approximate linear methods to be used (Joachims, 2006) for inference tasks. If $\mathbf{M} = \mathbf{G}$, the unstructured Gaussian matrix, then we obtain the standard random feature estimator. We shall contrast this approach against the use of matrices from the ROMs family.

When constructing random feature maps for kernels, very often $m > n$. In this case, our structured mechanism can be applied by concatenating some number of independent structured blocks. Our theoretical guarantees will be given just for one block, but can easily be extended to a larger number of blocks since different blocks are independent.

The standard random feature approximation $\widehat{K}_m^{\mathrm{ang,base}}$ for approximating the angular kernel is defined by taking $\mathbf{M}$ to be $\mathbf{G}$, the unstructured Gaussian matrix, in Equation (6), and satisfies the following.

Lemma 4.2. *The estimator $\widehat{K}_m^{\mathrm{ang,base}}$ is unbiased and $\mathrm{MSE}(\widehat{K}_m^{\mathrm{ang,base}}(\mathbf{x},\mathbf{y})) = \frac{4\theta_{\mathbf{x},\mathbf{y}}(\pi-\theta_{\mathbf{x},\mathbf{y}})}{m\pi^2}$.*

The MSE of an estimator $\widehat{K}^{\mathrm{ang}}(\mathbf{x},\mathbf{y})$ of the true angular kernel $K^{\mathrm{ang}}(\mathbf{x},\mathbf{y})$ is defined analogously to the MSE of an estimator of the dot product, given in §3. Our main result regarding angular kernels states that if we instead take $\mathbf{M} = \mathbf{G}_{\mathrm{ort}}$ in Equation (6), then we obtain an estimator $\widehat{K}_m^{\mathrm{ang,ort}}$ with strictly smaller MSE, as follows.

Theorem 4.3. *Estimator $\widehat{K}_m^{\mathrm{ang,ort}}$ is unbiased and satisfies:*

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{ang,ort}}(\mathbf{x},\mathbf{y})) < \mathrm{MSE}(\widehat{K}_m^{\mathrm{ang,base}}(\mathbf{x},\mathbf{y})).$$

We also derive a formula for the MSE of an estimator $\widehat{K}_m^{\mathrm{ang},\mathbf{M}}$ of the angular kernel which replaces $\mathbf{G}$ with an arbitrary random matrix $\mathbf{M}$ and uses m random feature maps. The formula is helpful to see how the quality of the estimator depends on the probabilities that the projections of the rows of $\mathbf{M}$ are contained in some particular convex regions of the 2-dimensional space $\mathcal{L}_{\mathbf{x},\mathbf{y}}$ spanned by datapoints $\mathbf{x}$ and $\mathbf{y}$. For an illustration of the geometric definitions introduced in this Section, see Figure 1. The formula depends on probabilities involving events $\mathcal{A}^i = \{\mathrm{sgn}((\mathbf{r}^i)^T\mathbf{x}) \neq \mathrm{sgn}((\mathbf{r}^i)^T\mathbf{y})\}$, where $\mathbf{r}^i$ stands for the i^{th} row of the structured matrix. Notice that $\mathcal{A}^i = \{\mathbf{r}^i_{proj} \in \mathcal{C}_{\mathbf{x},\mathbf{y}}\}$, where $\mathbf{r}^i_{proj}$ stands for the projection of $\mathbf{r}^i$ into $\mathcal{L}_{\mathbf{x},\mathbf{y}}$ and $\mathcal{C}_{\mathbf{x},\mathbf{y}}$ is the union of two cones in $\mathcal{L}_{\mathbf{x},\mathbf{y}}$, each of angle $\theta_{\mathbf{x},\mathbf{y}}$.

Theorem 4.4. *Estimator $\widehat{K}_m^{\mathrm{ang},\mathbf{M}}$ satisfies the following, where: $\delta_{i,j} = \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}^j] - \mathbb{P}[\mathcal{A}^i]\mathbb{P}[\mathcal{A}^j]$:*

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{ang},\mathbf{M}}(\mathbf{x},\mathbf{y})) = \frac{1}{m^2}\left[m - \sum_{i=1}^{m}(1-2\mathbb{P}[\mathcal{A}^i])^2\right] + \frac{4}{m^2}\left[\sum_{i=1}^{m}(\mathbb{P}[\mathcal{A}^i] - \frac{\theta_{\mathbf{x},\mathbf{y}}}{\pi})^2 + \sum_{i\neq j}\delta_{i,j}\right].$$

Note that probabilities $\mathbb{P}[\mathcal{A}^i]$ and $\delta_{i,j}$ depend on the choice of $\mathbf{M}$. It is easy to prove that for unstructured $\mathbf{G}$ and $\mathbf{G}_{\mathrm{ort}}$ we have: $\mathbb{P}[\mathcal{A}^i] = \frac{\theta_{\mathbf{x},\mathbf{y}}}{\pi}$. Further, from the independence of the rows of $\mathbf{G}$, $\delta_{i,j} = 0$ for $i \neq j$. For unstructured $\mathbf{G}$ we obtain Lemma 4.2. Interestingly, we see that to prove Theorem 4.3, it suffices to show $\delta_{i,j} < 0$, which is the approach we take (see Appendix). If we replace $\mathbf{G}$ with $\mathbf{M}^{(k)}_{\mathbf{SR}}$, then the expression $\epsilon = \mathbb{P}[\mathcal{A}^i] - \frac{\theta_{\mathbf{x},\mathbf{y}}}{\pi}$ does not depend on i. Hence, the angular kernel estimator based on Hadamard matrices gives smaller MSE estimator if and only if $\sum_{i\neq j}\delta_{i,j} + m\epsilon^2 < 0$. It is not yet clear if this holds in general.

As alluded to at the beginning of this section, the angular kernel may be viewed as a member of a wie family of kernels known as Pointwise Nonlinear Gaussian kernels.

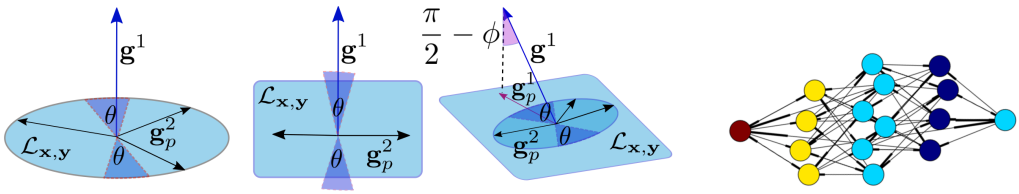


Figure 1: **Left part:** Left: $\mathbf{g}^1$ is orthogonal to $\mathcal{L}_{\mathbf{x},\mathbf{y}}$. Middle: $\mathbf{g}^1 \in \mathcal{L}_{\mathbf{x},\mathbf{y}}$. Right: $\mathbf{g}^1$ is close to orthogonal to $\mathcal{L}_{\mathbf{x},\mathbf{y}}$. **Right part:** Visualization of the Cayley graph explored by the Hadamard-Rademacher process in two dimensions. Nodes are colored red, yellow, light blue, dark blue, for Cayley distances of $0, 1, 2, 3$ from the identity matrix respectively. See text in §5.

Definition 4.5. *For a given function f, the Pointwise Nonlinear Gaussian kernel (PNG) K^f is defined by $K^f(\mathbf{x}, \mathbf{y}) = \mathbb{E}\left[f(\mathbf{g}^T\mathbf{x})f(\mathbf{g}^T\mathbf{y})\right]$, where $\mathbf{g}$ is a Gaussian vector with i.i.d $\mathcal{N}(0,1)$ entries.*

Many prominent examples of kernels (Williams, 1998; Cho and Saul, 2009) are PNGs. Wiener's tauberian theorem shows that all stationary kernels may be approximated arbitrarily well by sums of PNGs (Samo and Roberts, 2015). In future work we hope to explore whether ROMs can be used to achieve statistical benefit in estimation tasks associated with a wider range of PNGs.

5 Understanding the effectiveness of orthogonality

Here we build intuitive understanding for the effectiveness of ROMs. We examine geometrically the angular kernel (see §4), then discuss a connection to random walks over orthogonal matrices.

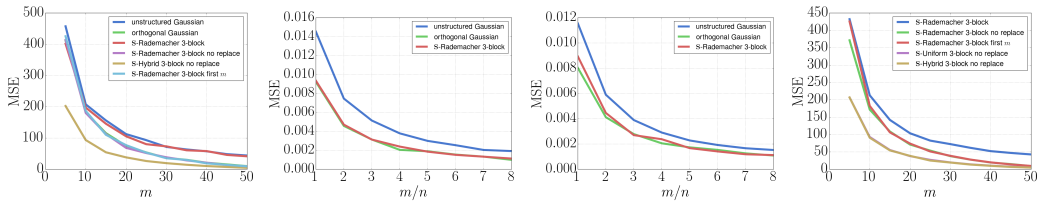
Angular kernel. As noted above for the $\mathbf{G}_{\mathrm{ort}}$-mechanism, smaller MSE than that for unstructured $\mathbf{G}$ is implied by the inequality $\mathbb{P}[\mathcal{A}^i \cap \mathcal{A}^j] < \mathbb{P}[\mathcal{A}^i]\mathbb{P}[\mathcal{A}^j]$, which is equivalent to: $\mathbb{P}[\mathcal{A}^j|\mathcal{A}^i] < \mathbb{P}[\mathcal{A}^j]$. Now it becomes clear why orthogonality is crucial. Without loss of generality take: $i = 1$, $j = 2$, and let $\mathbf{g}^1$ and $\mathbf{g}^2$ be the first two rows of $\mathbf{G}_{\mathrm{ort}}$.

Consider first the extreme case (middle of left part of Figure 1), where all vectors are 2-dimensional. Recall definitions from just after Theorem 4.3. If $\mathbf{g}^1$ is in $\mathcal{C}_{\mathbf{x},\mathbf{y}}$ then it is much less probable for $\mathbf{g}^2$ also to belong to $\mathcal{C}_{\mathbf{x},\mathbf{y}}$. In particular, if $\theta < \frac{\pi}{2}$ then the probability is zero. That implies the inequality. On the other hand, if $\mathbf{g}^1$ is perpendicular to $\mathcal{L}_{\mathbf{x},\mathbf{y}}$ then conditioning on $\mathcal{A}^i$ does not have any effect on the probability that $\mathbf{g}^2$ belongs to $\mathcal{C}_{\mathbf{x},\mathbf{y}}$ (left subfigure of Figure 1). In practice, with high probability the angle ϕ between $\mathbf{g}^1$ and $\mathcal{L}_{\mathbf{x},\mathbf{y}}$ is close to $\frac{\pi}{2}$, but is not exactly $\frac{\pi}{2}$. That again implies that conditioned on the projection $\mathbf{g}_p^1$ of $\mathbf{g}^1$ into $\mathcal{L}_{\mathbf{x},\mathbf{y}}$ to be in $\mathcal{C}_{\mathbf{x},\mathbf{y}}$, the more probable directions of $\mathbf{g}_p^2$ are perpendicular to $\mathbf{g}_p^1$ (see: ellipsoid-like shape in the right subfigure of Figure 1 which is the projection of the sphere taken from the $(n-1)$-dimensional space orthogonal to $\mathbf{g}^1$ into $\mathcal{L}_{\mathbf{x},\mathbf{y}}$). This makes it less probable for $\mathbf{g}_p^2$ to be also in $\mathcal{C}_{\mathbf{x},\mathbf{y}}$. The effect is subtle since $\phi \approx \frac{\pi}{2}$, but this is what provides superiority of the orthogonal transformations over state-of-the-art ones in the angular kernel approximation setting.

Markov chain perspective. We focus on Hadamard-Rademacher random matrices $\mathbf{HD}_k...\mathbf{HD}_1$, a special case of the $\mathbf{SD}$-product matrices described in Section 2. Our aim is to provide intuition for how the choice of k affects the quality of the random matrix, following our earlier observations just after Corollary 3.4, which indicated that for $\mathbf{SD}$-product matrices, odd values of k yield greater benefits than even values, and that there are diminishing benefits from higher values of k. We proceed by casting the random matrices into the framework of Markov chains.

Definition 5.1. *The Hadamard-Rademacher process in n dimensions is the Markov chain $(\mathbf{X}_k)_{k=0}^{\infty}$ taking values in the orthogonal group $O(n)$, with $\mathbf{X}_0 = \mathbf{I}$ almost surely, and $\mathbf{X}_k = \mathbf{HD}_k\mathbf{X}_{k-1}$ almost surely, where $\mathbf{H}$ is the normalized Hadamard matrix in n dimensions, and $(\mathbf{D}_k)_{k=1}^{\infty}$ are iid diagonal matrices with independent Rademacher random variables on their diagonals.*

Constructing an estimator based on Hadamard-Rademacher matrices is equivalent to simulating several time steps from the Hadamard-Rademacher process. The quality of estimators based on Hadamard-Rademacher random matrices comes from a quick mixing property of the corresponding

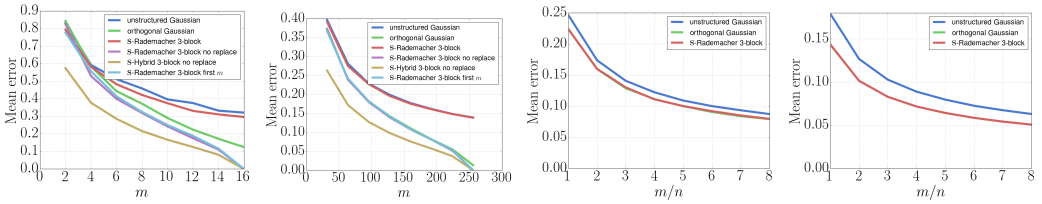


(a) `g50c` - pointwise evaluation MSE for inner product estimation

(b) `random` - angular kernel

(c) `random` - angular kernel with true angle $\pi/4$

(d) `g50c` - inner product estimation MSE for variants of 3-block **SD**-product matrices.



(e) `LETTER` - dot-product

(f) `USPS` - dot-product

(g) `LETTER` - angular kernel

(h) `USPS` - angular kernel

Figure 2: **Top row:** MSE curves for pointwise approximation of inner product and angular kernels on the `g50c` dataset, and randomly chosen vectors. **Bottom row:** Gram matrix approximation error for a variety of data sets, projection ranks, transforms, and kernels. Note that the error scaling is dependent on the application.

Markov chain. The following demonstrates attractive properties of the chain in low dimensions.

Proposition 5.2. *The Hadamard-Rademacher process in two dimensions: explores a state-space of* 16 *orthogonal matrices, is ergodic with respect to the uniform distribution on this set, has period* 2*, the diameter of the Cayley graph of its state space is* 3*, and the chain is fully mixed after* 3 *time steps.*

This proposition, and the Cayley graph corresponding to the Markov chain's state space (Figure 1 right), illustrate the fast mixing properties of the Hadamard-Rademacher process in low dimensions; this agrees with the observations in §3 that there are diminishing returns associated with using a large number k of **HD** blocks in an estimator. The observation in Proposition 5.2 that the Markov chain has period 2 indicates that we should expect different behavior for estimators based on odd and even numbers of blocks of **HD** matrices, which is reflected in the analytic expressions for MSE derived in Theorems 3.3 and 3.6 for the dimensionality reduction setup.

6 Experiments

We present comparisons of estimators introduced in §3 and §4, illustrating our theoretical results, and further demonstrating the empirical success of ROM-based estimators at the level of Gram matrix approximation. We compare estimators based on: unstructured Gaussian matrices $\mathbf{G}$, matrices $\mathbf{G}_{\mathrm{ort}}$, $\mathbf{S}$-*Rademacher* and $\mathbf{S}$-*Hybrid* matrices with $k = 3$ and different sub-sampling strategies. Results for $k > 3$ do not show additional statistical gains empirically. Additional experimental results, including a comparison of estimators using different numbers of **SD** blocks, are in the Appendix §10. Throughout, we use the normalized Hadamard matrix $\mathbf{H}$ for the structured matrix $\mathbf{S}$.

6.1 Pointwise kernel approximation

Complementing the theoretical results of §3 and §4, we provide several salient comparisons of the various methods introduced - see Figure 2 top. Plots presented here (and in the Appendix) compare MSE for dot-product and angular and kernel. They show that estimators based on $\mathbf{G}_{\mathrm{ort}}$, $\mathbf{S}$-*Hybrid* and $\mathbf{S}$-*Rademacher* matrices without replacement, or using the first m rows, beat the state-of-the-art unstructured $\mathbf{G}$ approach on accuracy for all our different datasets in the JLT setup. Interestingly, the latter two approaches give also smaller MSE than $\mathbf{G}_{\mathrm{ort}}$-estimators. For angular kernel estimation, where sampling is not relevant, we see that $\mathbf{G}_{\mathrm{ort}}$ and $\mathbf{S}$-*Rademacher* approaches again outperform the ones based on matrices $\mathbf{G}$.

6.2 Gram matrix approximation

Moving beyond the theoretical guarantees established in §3 and §4, we show empirically that the superiority of estimators based on ROMs is maintained at the level of Gram matrix approximation. We compute Gram matrix approximations (with respect to both standard dot-product, and angular kernel) for a variety of datasets. We use the normalized Frobenius norm error $\|\mathbf{K} - \widehat{\mathbf{K}}\|_2 / \|\mathbf{K}\|_2$ as our metric (as used by Choromanski and Sindhwani, 2016), and plot the mean error based on 1,000 repetitions of each random transform - see Figure 2 bottom. The Gram matrices are computed on a randomly selected subset of 550 data points from each dataset. As can be seen, the $\mathbf{S}$-*Hybrid* estimators using the "no-replacement" or "first m rows" sub-sampling strategies outperform even the orthogonal Gaussian ones in the dot-product case. For the angular case, the $\mathbf{G}_{\mathrm{ort}}$-approach and $\mathbf{S}$-*Rademacher* approach are practically indistinguishable.

7 Conclusion

We defined the family of random ortho-matrices (ROMs). This contains the $\mathbf{SD}$-product matrices, which include a number of recently proposed structured random matrices. We showed theoretically and empirically that ROMs have strong statistical and computational properties (in several cases outperforming previous state-of-the-art) for algorithms performing dimensionality reduction and random feature approximations of kernels. We highlight Corollary 3.4, which provides a theoretical guarantee that $\mathbf{SD}$-product matrices yield better accuracy than iid matrices in an important dimensionality reduction application (we believe the first result of this kind). Intriguingly, for dimensionality reduction, using just one complex structured matrix yields random features of much better quality. We provided perspectives to help understand the benefits of ROMs, and to help explain the behavior of $\mathbf{SD}$-product matrices for various numbers of blocks. Our empirical findings suggest that our theoretical results might be further strengthened, particularly in the kernel setting.

Acknowledgements

We thank Vikas Sindhwani at Google Brain Robotics and Tamas Sarlos at Google Research for inspiring conversations that led to this work. We thank Matej Balog, Maria Lomeli, Jiri Hron and Dave Janz for helpful comments. MR acknowledges support by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/L016516/1 for the University of Cambridge Centre for Doctoral Training, the Cambridge Centre for Analysis. AW acknowledges support by the Alan Turing Institute under the EPSRC grant EP/N510129/1, and by the Leverhulme Trust via the CFI.

References

N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *STOC*, 2006.

A. Andoni, P. Indyk, T. Laarhoven, I. Razenshteyn, and L. Schmidt. Practical and optimal LSH for angular distance. In *NIPS*, 2015.

M. Bojarski, A. Choromanska, K. Choromanski, F. Fagan, C. Gouy-Pailler, A. Morvan, N. Sakr, T. Sarlos, and J. Atif. Structured adaptive and random spinners for fast machine learning computations. In *to appear in AISTATS*, 2017.

Y. Cho and L. K. Saul. Kernel methods for deep learning. In *NIPS*, 2009.

A. Choromanska, K. Choromanski, M. Bojarski, T. Jebara, S. Kumar, and Y. LeCun. Binary embeddings with structured hashed projections. In *ICML*, 2016.

K. Choromanski and V. Sindhwani. Recycling randomness with structure for sublinear time kernel expansions. In *ICML*, 2016.

A. Hinrichs and J. Vybíral. Johnson-Lindenstrauss lemma for circulant matrices. *Random Structures & Algorithms*, 39(3):391–398, 2011.

H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.

Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 217–226, New York, NY, USA, 2006. ACM. ISBN 1-59593-339-5. doi: 10.1145/1150402.1150429. URL `http://doi.acm.org/10.1145/1150402.1150429`.

W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Q. Le, T. Sarlós, and A. Smola. Fastfood - approximating kernel expansions in loglinear time. In *ICML*, 2013.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, 2007.

Y.-L. K. Samo and S. Roberts. Generalized spectral kernels. *CoRR*, abs/1506.02236, 2015.

L. Schmidt, M. Sharifi, and I. Moreno. Large-scale speaker identification. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1650–1654. IEEE, 2014.

G. Sidorov, A. Gelbukh, H. Gómez-Adorno, and D. Pinto. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3), 2014.

N. Sundaram, A. Turmukhametova, N. Satish, T. Mostak, P. Indyk, S. Madden, and P. Dubey. Streaming similarity search over one billion tweets using parallel locality-sensitive hashing. *Proceedings of the VLDB Endowment*, 6(14):1930–1941, 2013.

J. Vybíral. A variant of the Johnson-Lindenstrauss lemma for circulant matrices. *Journal of Functional Analysis*, 260(4):1096–1105, 2011.

C. Williams. Computation with infinite neural networks. *Neural Computation*, 10(5):1203–1216, 1998.

F. Yu, A. Suresh, K. Choromanski, D. Holtmann-Rice, and S. Kumar. Orthogonal random features. In *NIPS*, pages 1975–1983, 2016.

H. Zhang and L. Cheng. New bounds for circulant Johnson-Lindenstrauss embeddings. *CoRR*, abs/1308.6339, 2013.

Xu Zhang, Felix X. Yu, Ruiqi Guo, Sanjiv Kumar, Shengjin Wang, and Shih-Fu Chang. Fast orthogonal projection based on kronecker product. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2929–2937, 2015. doi: 10.1109/ICCV.2015.335. URL `http://dx.doi.org/10.1109/ICCV.2015.335`.

APPENDIX:
The Unreasonable Effectiveness of Random Orthogonal Embeddings

We present here details and proofs of all the theoretical results presented in the main body of the paper. We also provide further experimental results in §10.

We highlight proofs of several key results that may be of particular interest to the reader:

- The proof of Theorem 3.3; see §8.3.
- The proof of Theorem 3.6; see §8.5.
- The proof of Theorem 4.3; see §9.2.

In the Appendix we will use interchangeably two notations for the dot product between vectors $\mathbf{x}$ and $\mathbf{y}$, namely: $\mathbf{x}^\top \mathbf{y}$ and $\langle \mathbf{x}, \mathbf{y} \rangle$.

8 Proofs of results in §3

8.1 Proof of Lemma 3.1

Proof. Denote $X_i = (\mathbf{g}^i)^\top \mathbf{x} \cdot (\mathbf{g}^i)^\top \mathbf{y}$, where $\mathbf{g}^i$ stands for the i^{th} row of the unstructured Gaussian matrix $\mathbf{G} \in \mathbb{R}^{m \times n}$. Note that we have:

$$\widehat{K}_m^{\mathrm{base}}(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{i=1}^{m} X_i. \tag{7}$$

Denote $\mathbf{g}^i = (g_1^i, ..., g_n^i)^\top$. Notice that from the independence of g_j^is and the fact that: $\mathbb{E}[g_j^i] = 0$, $\mathbb{E}[(g_j^i)^2] = 1$, we get: $\mathbb{E}[X_i] = \sum_{i=1}^n x_i y_i = \mathbf{x}^\top \mathbf{y}$, thus the estimator is unbiased. Since the estimator is unbiased, we have: $\mathrm{MSE}(\widehat{K}_m^{\mathrm{base}}(\mathbf{x}, \mathbf{y})) = Var(\widehat{K}_m^{\mathrm{base}}(\mathbf{x}, \mathbf{y}))$. Thus we get:

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{base}}(\mathbf{x}, \mathbf{y})) = \frac{1}{m^2} \sum_{i,j} (\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]). \tag{8}$$

From the independence of different X_is, we get:

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{base}}(\mathbf{x}, \mathbf{y})) = \frac{1}{m^2} \sum_{i} (\mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2). \tag{9}$$

Now notice that different X_is have the same distribution, thus we get:

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{base}}(\mathbf{x}, \mathbf{y})) = \frac{1}{m} (\mathbb{E}[X_1^2] - (\mathbb{E}[X_1])^2). \tag{10}$$

From the unbiasedness of the estimator, we have: $\mathbb{E}[X_1] = \mathbf{x}^\top \mathbf{y}$. Therefore we obtain:

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{base}}(\mathbf{x}, \mathbf{y})) = \frac{1}{m} (\mathbb{E}[X_1^2] - (\mathbf{x}^\top \mathbf{y})^2). \tag{11}$$

Now notice that

$$\mathbb{E}[X_1^2] = \mathbb{E}[\sum_{i_1, j_1, i_2, j_2} g_{i_1} g_{j_1} g_{i_2} g_{j_2} x_{i_1} y_{j_1} x_{i_2} y_{j_2}] = \sum_{i_1, j_1, i_2, j_2} x_{i_1} y_{j_1} x_{i_2} y_{j_2} \mathbb{E}[g_{i_1} g_{j_1} g_{i_2} g_{j_2}], \tag{12}$$

where $(g_1, ..., g_n)$ stands for the first row of $\mathbf{G}$. In the expression above the only nonzero terms corresponds to quadruples (i_1, j_1, i_2, j_2), where no index appears odd number of times. Therefore, from the inclusion-exclusion principle and the fact that $\mathbb{E}[g_i^2] = 1$ and $\mathbb{E}[g_i^4] = 3$, we obtain

$$\mathbb{E}[X_1^2] = \sum_{i_1 = j_1, i_2 = j_2} x_{i_1} y_{j_1} x_{i_2} y_{j_2} \mathbb{E}[g_{i_1} g_{j_1} g_{i_2} g_{j_2}] + \sum_{i_1 = i_2, j_1 = j_2} x_{i_1} y_{j_1} x_{i_2} y_{j_2} \mathbb{E}[g_{i_1} g_{j_1} g_{i_2} g_{j_2}] \tag{13}$$

$$+\sum_{i_1=j_2,i_2=j_1} x_{i_1}y_{j_1}x_{i_2}y_{j_2}\mathbb{E}[g_{i_1}g_{j_1}g_{i_2}g_{j_2}] - \sum_{i_1=j_1=i_2=j_2} x_{i_1}y_{j_1}x_{i_2}y_{j_2}\mathbb{E}[g_{i_1}g_{j_1}g_{i_2}g_{j_2}] \tag{14}$$

$$= ((\mathbf{x}^\top\mathbf{y})^2 - \sum_{i=1}^n x_i^2y_i^2 + 3\sum_{i=1}^n x_i^2y_i^2) + ((\|\mathbf{x}\|_2\|\mathbf{y}\|_2)^2 - \sum_{i=1}^n x_i^2y_i^2 + 3\sum_{i=1}^n x_i^2y_i^2) \tag{15}$$

$$+ ((\mathbf{x}^\top\mathbf{y})^2 - \sum_{i=1}^n x_i^2y_i^2 + 3\sum_{i=1}^n x_i^2y_i^2) - 3\cdot 2\sum_{i=1}^n x_i^2y_i^2 \tag{16}$$

$$= (\|\mathbf{x}\|_2\|\mathbf{y}\|_2)^2 + 2(\mathbf{x}^\top\mathbf{y})^2. \tag{17}$$

Therefore we obtain

$$\mathrm{MSE}(\widehat{K}_m^{\mathrm{base}}(\mathbf{x},\mathbf{y})) = \frac{1}{m}((\|\mathbf{x}\|_2\|\mathbf{y}\|_2)^2 + 2(\mathbf{x}^\top\mathbf{y})^2 - (\mathbf{x}^\top\mathbf{y})^2) = \frac{1}{m}(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + (\mathbf{x}^\top\mathbf{y})^2), \tag{18}$$

which completes the proof. □

8.2 Proof of Theorem 3.2

Proof. The unbiasedness of the Gaussian orthogonal estimator comes from the fact that every row of the Gaussian orthogonal matrix is sampled from multivariate Gaussian distribution with entries taken independently at random from $\mathcal{N}(0,1)$.

Note that:

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\mathbb{E}[X_j], \tag{19}$$

where: $X_i = (\mathbf{r}_i^\top\mathbf{x})(\mathbf{r}_i^\top\mathbf{y})$, $X_j = (\mathbf{r}_j^\top\mathbf{x})(\mathbf{r}_j^\top\mathbf{y})$ and $\mathbf{r}_i, \mathbf{r}_j$ stand for the i^{th} and j^{th} row of the Gaussian orthogonal matrix respectively. From the fact that Gaussian orthogonal estimator is unbiased, we get:

$$\mathbb{E}[X_i] = \mathbf{x}^\top\mathbf{y}. \tag{20}$$

Let us now compute $\mathbb{E}[X_iX_j]$. Writing $\mathbf{Z}_1 = \mathbf{r}_i$, $\mathbf{Z}_2 = \mathbf{r}_j$, we begin with some geometric observations:

- If $\phi \in [0, \pi/2]$ is the acute angle between $\mathbf{Z}_1$ and the $\mathbf{x}$-$\mathbf{y}$ plane, then ϕ has density $f(\phi) = (n-2)\cos(\phi)\sin^{n-3}(\phi)$.
- The squared norm of the projection of $\mathbf{Z}_1$ into the $\mathbf{x}$-$\mathbf{y}$ plane is therefore given by the product of a χ_n^2 random variable (the norm of $\mathbf{Z}_2$), multiplied by $\cos^2(\phi)$, where ϕ is distributed as described above, independently from the χ_n^2 random variable.
- The angle $\psi \in [0, 2\pi)$ between $\mathbf{x}$ and the projection of $\mathbf{Z}_1$ into the $\mathbf{x}$-$\mathbf{y}$ plane is distributed uniformly.
- Conditioned on the angle ϕ, the direction of $\mathbf{Z}_2$ is distributed uniformly on the hyperplane of $\mathbb{R}^n$ orthogonal to $\mathbf{Z}_1$. Using hyperspherical coordinates for the unit hypersphere of this hyperplane, we may pick an orthonormal basis of the $\mathbf{x}$-$\mathbf{y}$ plane such that the first basis vector is the unit vector in the direction of the projection of $\mathbf{Z}_1$, and the coordinates of the projection of $\mathbf{Z}_2$ with respect to this basis are $(\sin(\phi)\cos(\varphi_1), \sin(\varphi_1)\cos(\varphi_2))$, where φ_1, φ_2 are random angles taking values in $[0, \pi]$, with densities given by $\sin^{n-3}(\varphi_1)I(n-3)^{-1}$ and $\sin^{n-4}(\varphi_2)I(n-4)^{-1}$ respectively. Here $I(k) = \int_0^\pi \sin^k(x)dx = \sqrt{\pi}\Gamma((k+1)/2)/\Gamma(k/2+1)$.
- The angle t that the projection of $\mathbf{Z}_2$ into the $\mathbf{x}$-$\mathbf{y}$ plane makes with the projection of $\mathbf{Z}_1$ then satisfies $\tan(t) = \sin(\varphi_1)\cos(\varphi_2)/(\sin(\phi)\cos(\varphi_1)) = \cos(\varphi_1)/\sin(\phi) \times \tan(\varphi_1)$.

Applying these observations, we get:

$$\begin{aligned}
&\mathbb{E}[X_iX_j] \\
=&\mathbb{E}[(\mathbf{r}_i^\top\mathbf{x})(\mathbf{r}_i^\top\mathbf{y})(\mathbf{r}_j^\top\mathbf{x})(\mathbf{r}_j^\top\mathbf{y})] \\
=&\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 n^2 \int_0^{\pi/2} d\phi f(\phi)\cos^2(\phi) \int_0^\pi d\varphi_1 \sin^{n-3}(\varphi_1)I(n-3)^{-1} \int_0^\pi d\varphi_2 \sin^{n-4}(\varphi_2)I(n-4)^{-1}\times
\end{aligned}$$

$$\int_0^{2\pi} \frac{d\psi}{2\pi} \left(\sin^2(\phi)\cos^2(\varphi_1) + \sin^2(\varphi_1)\cos^2(\varphi_2)\right) \cos(\psi)\cos(\psi+\theta)\cos(t-\psi)\cos(t-\theta-\psi). \tag{21}$$

We first apply the cosine product formula to the two adjacent pairs making up the final product of four cosines involving ψ in the integrand above. The majority of these terms vanish upon integrating with respect to ψ, due to the periodicity of the integrands wrt ψ. We are thus left with:

$$\begin{aligned}
&\mathbb{E}[X_i X_j] \\
=&\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 n^2 \int_0^{\pi/2} d\phi f(\phi)\cos^2(\phi) \int_0^{\pi} d\varphi_1 \sin^{n-3}(\varphi_1) I(n-3)^{-1} \int_0^{\pi} d\varphi_2 \sin^{n-4}(\varphi_2) I(n-4)^{-1} \times \\
&\left(\sin^2(\phi)\cos^2(\varphi_1) + \sin^2(\varphi_1)\cos^2(\varphi_2)\right) \left(\frac{1}{4}\cos^2(\theta) + \frac{1}{8}\cos(2t)\right).
\end{aligned} \tag{22}$$

We now consider two constituent parts of the integral above: one involving the term $\frac{1}{4}\cos^2(\theta)$, and the other involving $\frac{1}{8}\cos(2t)$. We deal first with the former; its evaluation requires several standard trigonometric integrals:

$$\begin{aligned}
&\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 n^2 \int_0^{\pi/2} d\phi f(\phi)\cos^2(\phi) \int_0^{\pi} d\varphi_1 \sin^{n-3}(\varphi_1) I(n-3)^{-1} \int_0^{\pi} d\varphi_2 \sin^{n-4}(\varphi_2) I(n-4)^{-1} \times \\
&\left(\sin^2(\phi)\cos^2(\varphi_1) + \sin^2(\varphi_1)\cos^2(\varphi_2)\right) \frac{1}{4}\cos^2(\theta) \\
=&\frac{\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 n^2 \cos^2(\theta)}{4I(n-3)I(n-4)} \int_0^{\pi/2} d\phi f(\phi)\cos^2(\phi) \int_0^{\pi} d\varphi_1 \sin^{n-3}(\varphi_1) \times \\
&\qquad \left(\sin^2(\phi)\cos^2(\varphi_1) I(n-4) + \sin^2(\varphi_1)\left(I(n-4) - I(n-2)\right)\right) \\
=&\frac{\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 n^2 \cos^2(\theta)}{4I(n-3)I(n-4)} \int_0^{\pi/2} d\phi (n-2)\sin^{n-3}(\phi)\cos(\phi)\cos^2(\phi) \times \\
&\qquad \left(\sin^2(\phi)(I(n-3) - I(n-1))I(n-4) + I(n-1)\left(I(n-4) - I(n-2)\right)\right) \\
=&\frac{\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 n^2 \cos^2(\theta)}{4I(n-3)I(n-4)} \left(\left(\frac{1}{n} - \frac{1}{n+2}\right)(I(n-3) - I(n-1))I(n-4) + \right. \\
&\qquad \left. I(n-1)\left(I(n-4) - I(n-2)\right)\left(\frac{1}{n-2} - \frac{1}{n}\right)\right).
\end{aligned} \tag{23}$$

We may now turn our attention to the other constituent integral of Equation (22), which involves the term $\cos(2t)$. Recall that from our earlier geometric considerations, we have $\tan(t) = \frac{\cos(\varphi_2)}{\sin(\phi)}\tan(\phi_1)$. An elementary trigonometric calculation using the tan half-angle formula yields:

$$\begin{aligned}
\cos(2t) &= \cos\left(2\arctan\left(\frac{\cos(\varphi_2)}{\sin(\phi)}\tan(\varphi_1)\right)\right) \\
&= \frac{1 - \frac{\cos^2(\varphi_2)}{\sin^2(\phi)}\tan^2(\varphi_1)}{\frac{\cos^2(\varphi_2)}{\sin^2(\phi)}\tan^2(\varphi_1) + 1} \\
&= \frac{\sin^2(\phi)\cos^2(\varphi_1) - \cos^2(\varphi_2)\sin^2(\varphi_1)}{\cos^2(\varphi_2)\sin^2(\varphi_1) + \sin^2(\phi)\cos^2(\varphi_1)}.
\end{aligned} \tag{24}$$

This observation greatly simplifies the integral from Equation (22) involving the term $\cos(2t)$, as follows:

$$\begin{aligned}
&\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 n^2 \int_0^{\pi/2} d\phi f(\phi)\cos^2(\phi) \int_0^{\pi} d\varphi_1 \sin^{n-3}(\varphi_1) I(n-3)^{-1} \int_0^{\pi} d\varphi_2 \sin^{n-4}(\varphi_2) I(n-4)^{-1} \times \\
&\left(\sin^2(\phi)\cos^2(\varphi_1) + \sin^2(\varphi_1)\cos^2(\varphi_2)\right) \frac{1}{8}\cos(2t) \\
=&\frac{\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 n^2}{8I(n-3)I(n-4)} \int_0^{\pi/2} d\phi f(\phi)\cos^2(\phi) \int_0^{\pi} d\varphi_1 \sin^{n-3}(\varphi_1) \int_0^{\pi} d\varphi_2 \sin^{n-4}(\varphi_2) \times
\end{aligned}$$

$$\left(\sin^2(\phi)\cos^2(\varphi_1)+\sin^2(\varphi_1)\cos^2(\varphi_2)\right)\frac{\sin^2(\phi)\cos^2(\varphi_1)-\cos^2(\varphi_2)\sin^2(\varphi_1)}{\cos^2(\varphi_2)\sin^2(\varphi_1)+\sin^2(\phi)\cos^2(\varphi_1)}$$
$$=\frac{\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2n^2}{8I(n-3)I(n-4)}\int_0^{\pi/2}d\phi f(\phi)\cos^2(\phi)\int_0^{\pi}d\varphi_1\sin^{n-3}(\varphi_1)\int_0^{\pi}d\varphi_2\sin^{n-4}(\varphi_2)\times$$
$$\left(\sin^2(\phi)\cos^2(\varphi_1)-\cos^2(\varphi_2)\sin^2(\varphi_1)\right). \tag{25}$$

But now observe that this integral is exactly of the form dealt with in (23), hence we may immediately identify its value as:

$$\frac{\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2n^2}{8I(n-3)I(n-4)}\Bigg(\left(\frac{1}{n}-\frac{1}{n+2}\right)(I(n-3)-I(n-1))I(n-4)-$$
$$I(n-1)\left(I(n-4)-I(n-2)\right)\left(\frac{1}{n-2}-\frac{1}{n}\right)\Bigg). \tag{26}$$

Thus substituting our calculations back into Equation (22), we obtain:

$$\mathbb{E}[X_iX_j]$$
$$=\frac{\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2n^2}{4I(n-3)I(n-4)}\Bigg(\left(\frac{1}{n}-\frac{1}{n+2}\right)(I(n-3)-I(n-1))I(n-4)\left[\cos^2(\theta)+\frac{1}{2}\right]+$$
$$I(n-1)\left(I(n-4)-I(n-2)\right)\left(\frac{1}{n-2}-\frac{1}{n}\right)\left[\cos^2(\theta)-\frac{1}{2}\right]\Bigg). \tag{27}$$

The covariance term is obtained by subtracting off $\mathbb{E}[X_i]\mathbb{E}[X_i]=\langle\mathbf{x},\mathbf{y}\rangle^2$. Now we sum over $m(m-1)$ covariance terms and take into account the normalization factor $\frac{1}{\sqrt{m}}$ for the Gaussian matrix entries. That gives the extra multiplicative term $\frac{m(m-1)}{m^2}=\frac{m-1}{m}$. Thus we obtain the quantity in the statement of the theorem, completing the proof. □

8.3 Proof of Theorem 3.3

We obtain Theorem 3.3 through a sequence of smaller propositions. Broadly, the strategy is first to show that the estimators of Theorem 3.3 are unbiased (Proposition 8.1). An expression for the mean squared error of the estimator $\widehat{K}_m^{(1)}$ with one matrix block is then derived (Proposition 8.2). Finally, a straightforward recursive formula for the mean squared error of the general estimator is derived (Proposition 8.3), and the result of the theorem then follows.

Proposition 8.1. *The estimator $\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y})$ is unbiased, for all $k,n\in\mathbb{N}$, $m\leq n$, and $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$.*

Proof. Notice first that since rows of $\mathbf{S}=\{s_{i,j}\}$ are orthogonal and are L_2-normalized, the matrix $\mathbf{S}$ is an isometry. Thus each block $\mathbf{SD}_i$ is also an isometry. Therefore it suffices to prove the claim for $k=1$.

Then, denoting by $\mathbf{J}=(J_1,\ldots,J_m)$ the indices of the randomly selected rows of $\mathbf{SD}_1$, note that the estimator $\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})$ may be expressed in the form

$$\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})=\frac{1}{m}\sum_{i=1}^{m}\left(\sqrt{n}(\mathbf{SD}_1)_{J_i}\mathbf{x}\times\sqrt{n}(\mathbf{SD}_1)_{J_i}\mathbf{y}\right),$$

where $(\mathbf{SD}_1)_i$ is the i^{th} row of $\mathbf{SD}_1$. Since each of the rows of $\mathbf{SD}_1$ has the same marginal distribution, it suffices to demonstrate that $\mathbb{E}[\mathbf{y}^T\mathbf{D}_1\mathbf{S}_1^\top\mathbf{S}_1\mathbf{D}_1\mathbf{x}]=\frac{\mathbf{x}^\top\mathbf{y}}{n}$, where $\mathbf{S}_1$ is the first row of $\mathbf{S}$. Now note

$$\mathbb{E}[\mathbf{y}^\top\mathbf{D}\mathbf{S}_1^\top\mathbf{S}_1\mathbf{D}\mathbf{x}]=\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}y_id_i\times\sum_{i=1}^{n}x_id_i\right]=\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}x_iy_id_i^2\right]+\mathbb{E}\left[\sum_{i\neq j}x_iy_jd_id_j\right]=\frac{\mathbf{x}^\top\mathbf{y}}{n},$$

where $d_i=\mathbf{D}_{ii}$ are iid Rademacher random variables, for $i=1,\ldots,n$. □

With Proposition 8.1 in place, the mean square error for the estimator $\widehat{K}_m^{(1)}$ using one matrix block can be derived.

Proposition 8.2. *The MSE of the single $\mathbf{SD}^{(\mathcal{R})}$-block m-feature estimator $\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})$ for $\langle \mathbf{x},\mathbf{y}\rangle$ using the without replacement row sub-sampling strategy is*

$$\mathrm{MSE}(\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})) = \frac{1}{m}\left(\frac{n-m}{n-1}\right)\left(\|\mathbf{x}\|^2\|\mathbf{y}\|^2 + \langle \mathbf{x},\mathbf{y}\rangle^2 - 2\sum_{i=1}^{n} x_i^2 y_i^2\right).$$

Proof. First note that since $\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})$ is unbiased, the mean squared error is simply the variance of this estimator. Secondly, denoting the indices of the m randomly selected rows by $\mathbf{J} = (J_1,\ldots,J_m)$, by conditioning on $\mathbf{J}$ we obtain the following:

$$\mathrm{Var}\left(\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})\right) = \\ \frac{n^2}{m^2}\left(\mathbb{E}\left[\mathrm{Var}\left(\sum_{p=1}^{m}(\mathbf{SDx})_{J_p}(\mathbf{SDy})_{J_p}\middle|\mathbf{J}\right)\right] + \mathrm{Var}\left(\mathbb{E}\left[\sum_{p=1}^{m}(\mathbf{SDx})_{J_p}(\mathbf{SDy})_{J_p}\middle|\mathbf{J}\right]\right)\right).$$

Now note that the conditional expectation in the second term is constant as a function of J, since conditional on whichever rows are sampled, the resulting estimator is unbiased. Taking the variance of this constant therefore causes the second term to vanish. Now consider the conditional variance that appears in the first term:

$$\begin{aligned}\mathrm{Var}\left(\sum_{p=1}^{m}(\mathbf{SDx})_{J_p}(\mathbf{SDy})_{J_p}\middle|\mathbf{J}\right) &= \sum_{p=1}^{m}\sum_{p'=1}^{m}\mathrm{Cov}\left((\mathbf{SDx})_{J_m}(\mathbf{SDy})_{J_p},(\mathbf{SDx})_{J_{p'}}(\mathbf{SDy})_{J_{p'}}\middle|\mathbf{J}\right) \\ &= \sum_{p,p'=1}^{m}\sum_{i,j,k,l=1}^{n} s_{J_p i}s_{J_p j}s_{J_{p'}k}s_{J_{p'}l}x_i y_j x_k y_l \mathrm{Cov}\left(d_i d_j, d_k d_l\right),\end{aligned}$$

where we write $\mathbf{D} = \mathrm{Diag}(d_1,\ldots,d_n)$. Now note that $\mathrm{Cov}\left(d_i d_j, d_k d_l\right)$ is non-zero iff i,j are distinct, and $\{i,j\} = \{k,l\}$, in which case the covariance is 1. We therefore obtain:

$$\mathrm{Var}\left(\sum_{p=1}^{m}(\mathbf{SDx})_{J_p}(\mathbf{SDy})_{J_p}\middle|\mathbf{J}\right) = \\ \sum_{p,p'=1}^{m}\sum_{i\neq j}^{n}\left(s_{J_p i}s_{J_p j}s_{J_{p'}i}s_{J_{p'}j}x_i^2 y_j^2 + s_{J_p i}s_{J_p j}s_{J_{p'}j}s_{J_{p'}i}x_i y_j x_j y_i\right).$$

Substituting this expression for the conditional variance into the decomposition of the MSE of the estimator, we obtain the result of the theorem:

$$\begin{aligned}\mathrm{Var}\left(\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y})\right) &= \frac{n^2}{m^2}\mathbb{E}\left[\sum_{p,p'=1}^{m}\sum_{i\neq j}^{n}\left(s_{J_p i}s_{J_p j}s_{J_{p'}i}s_{J_{p'}j}x_i^2 y_j^2 + s_{J_p i}s_{J_p j}s_{J_{p'}j}s_{J_{p'}i}x_i y_j x_j y_i\right)\right] \\ &= \frac{n^2}{m^2}\sum_{p,p'=1}^{m}\sum_{i\neq j}^{n}\left(x_i^2 y_j^2 + x_i x_j y_i y_j\right)\mathbb{E}\left[s_{J_p i}s_{J_p j}s_{J_{p'}i}s_{J_{p'}j}\right].\end{aligned}$$

We now consider the law on the index variables $\mathbf{J} = (J_1,\ldots,J_m)$ induced by the sub-sampling strategy without replacement to evaluate the expectation in this last term. If $p = p'$, the integrand of the expectation is deterministically $1/n^2$. If $p \neq p'$, then we obtain:

$$\begin{aligned}\mathbb{E}\left[s_{J_p i}s_{J_p j}s_{J_{p'}i}s_{J_{p'}j}\right] &= \mathbb{E}\left[s_{J_p i}s_{J_p j}\mathbb{E}\left[s_{J_{p'}i}s_{J_{p'}j}\middle|J_p\right]\right] \\ &= \mathbb{E}\Bigg[s_{J_p i}s_{J_p j}\Bigg[\left(\frac{1}{n}\left(\frac{n/2-1}{n-1}\right) - \frac{1}{n}\left(\frac{n/2}{n-1}\right)\right)\mathbb{1}_{\{s_{J_p i}s_{J_p j}=1/n\}} + \\ &\qquad \left(\frac{1}{n}\left(\frac{n/2}{n-1}\right) - \frac{1}{n}\left(\frac{n/2-1}{n-1}\right)\right)\mathbb{1}_{\{s_{J_p i}s_{J_p j}=-1/n\}}\Bigg]\Bigg]\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n(n-1)} \mathbb{E}\left[s_{J_p i} s_{J_p j} \left(\mathbb{1}_{\{s_{J_p i} s_{J_p j} = -1/n\}} - \mathbb{1}_{\{s_{J_p i} s_{J_p j} = 1/n\}} \right) \right] \\
&= \frac{1}{n^2(n-1)} ,
\end{aligned}
$$

where we have used the fact that the products $s_{J_p i} s_{J_p j}$ and $s_{J_{p'} i} s_{J_{p'} j}$ take values in $\{\pm 1/n\}$, and because distinct rows of $\mathbf{S}$ are orthogonal, the marginal probability of each of the two values is $1/2$. A simple adjustment, using almost-sure distinctness of J_p and $J_{p'}$, yields the conditional probabilities needed to evaluate the conditional expectation that appears in the calculation above.

Substituting the values of these expectations back into the expression for the variance of $\widehat{K}_m^{(1)}(\mathbf{x}, \mathbf{y})$ then yields

$$
\begin{aligned}
\mathrm{Var}(\widehat{K}_m^{(1)}(\mathbf{x}, \mathbf{y})) &= \frac{n^2}{m^2} \sum_{i \neq j}^{n} \left(x_i^2 y_j^2 + x_i x_j y_i y_j \right) \left(m \times \frac{1}{n^2} - m(m-1) \times \frac{1}{n^2(n-1)} \right) \\
&= \frac{1}{m} \left(1 - \frac{m-1}{n-1} \right) \sum_{i \neq j}^{n} \left(x_i^2 y_j^2 + x_i x_j y_i y_j \right) \\
&= \frac{1}{m} \left(1 - \frac{m-1}{n-1} \right) \left(\sum_{i,j=1}^{n} (x_i^2 y_j^2 + x_i x_j y_i y_j) - 2 \sum_{i=1}^{n} x_i^2 y_i^2 \right) \\
&= \frac{1}{m} \left(\frac{n-m}{n-1} \right) \left(\langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 - 2 \sum_{i=1}^{n} x_i^2 y_i^2 \right) ,
\end{aligned}
$$

as required.

□

We now turn our attention to the following recursive expression for the mean squared error of a general estimator.

Proposition 8.3. *Let $k \geq 2$. We have the following recursion for the MSE of $K_m^{(k)}(x, y)$:*

$$
\mathrm{MSE}(\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y})) = \mathbb{E}\left[\mathrm{MSE}\left(\widehat{K}_m^{(k-1)}(\mathbf{S}\mathbf{D}_1\mathbf{x}, \mathbf{S}\mathbf{D}_1\mathbf{y}) | \mathbf{D}_1 \right) \right] .
$$

Proof. The result follows from a straightforward application of the law of total variance, conditioning on the matrix $\mathbf{D}_1$. Observe that

$$
\begin{aligned}
\mathrm{MSE}(\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y})) &= \mathrm{Var}(\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y})) \\
&= \mathbb{E}\left[\mathrm{Var}\left(\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y}) \middle| \mathbf{D}_1 \right) \right] + \mathrm{Var}\left(\mathbb{E}\left[\widehat{K}_m^{(k)}(\mathbf{x}, \mathbf{y}) \middle| \mathbf{D}_1 \right] \right) \\
&= \mathbb{E}\left[\mathrm{Var}\left(\widehat{K}_m^{(k-1)}(\mathbf{S}\mathbf{D}_1\mathbf{x}, \mathbf{S}\mathbf{D}_1\mathbf{y}) \middle| \mathbf{D}_1 \right) \right] + \mathrm{Var}\left(\mathbb{E}\left[\widehat{K}_m^{(k-1)}(\mathbf{S}\mathbf{D}_1\mathbf{x}, \mathbf{S}\mathbf{D}_1\mathbf{y}) \middle| \mathbf{D}_1 \right] \right) .
\end{aligned}
$$

But examining the conditional expectation in the second term, we observe

$$
\mathbb{E}\left[\widehat{K}_m^{(k-1)}(\mathbf{S}\mathbf{D}_1\mathbf{x}, \mathbf{S}\mathbf{D}_1\mathbf{y}) \middle| \mathbf{D}_1 \right] = \langle \mathbf{S}\mathbf{D}_1\mathbf{x}, \mathbf{S}\mathbf{D}_1\mathbf{y} \rangle \quad \text{almost surely} ,
$$

by unbiasedness of the estimator, and since $\mathbf{S}\mathbf{D}_1$ is orthogonal almost surely, this is equal to the (constant) inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ almost surely. This conditional expectation therefore has 0 variance, and so the second term in the expression for the MSE above vanishes, which results in the statement of the proposition. □

With these intermediate propositions established, we are now in a position to prove Theorem 3.3. In order to use the recursive result of Proposition 8.3, we require the following lemma.

Lemma 8.4. *For all $x, y, \in \mathbb{R}^n$, we have*

$$
\mathbb{E}\left[\sum_{i=1}^{n} (\mathbf{S}\mathbf{D}\mathbf{x})_i^2 (\mathbf{S}\mathbf{D}\mathbf{y})_i^2 \right] = \frac{1}{n} \left(\|\mathbf{x}\|^2 \|\mathbf{y}\|^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle^2 - 2 \sum_{i=1}^{n} x_i^2 y_i^2 \right) .
$$

Proof. The result follows by direct calculation. Note that

$$\mathbb{E}\left[\sum_{i=1}^{n}(\mathbf{SDx})_i^2(\mathbf{SDy})_i^2\right] = n\mathbb{E}\left[\left(\sum_{a=1}^{n}s_{1a}d_ax_a\right)^2\left(\sum_{a=1}s_{1a}d_ay_a\right)^2\right]$$
$$= n\sum_{i,j,k,l=1}^{n}s_{1i}s_{1j}s_{1k}s_{1l}x_ix_jy_ky_l\mathbb{E}\left[d_id_jd_kd_l\right],$$

where the first inequality follows since the n summands indexed by i in the initial expectation are identically distributed. Now note that the expectation $\mathbb{E}\left[d_id_jd_kd_l\right]$ is non-zero iff $i=j=k=l$, or $i=j\neq k=l$, or $i=k\neq j=l$, or $i=l\neq k=l$; in all such cases, the expectation takes the value 1. Substituting this into the above expression and collecting terms, we obtain

$$\mathbb{E}\left[\sum_{i=1}^{n}(\mathbf{SDx})_i^2(\mathbf{SDy})_i^2\right] = \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2y_i^2+\sum_{i\neq j}x_i^2y_i^2+2\sum_{i\neq j}x_ix_jy_iy_j\right)$$
$$= \frac{1}{n}\left(\sum_{i,j=1}^{n}x_i^2y_j^2+2\sum_{i,j=1}^{n}x_ix_jy_iy_j-2\sum_{i=1}^{n}x_i^2y_i^2\right),$$

from which the statement of the lemma follows immediately. □

Proof of Theorem 3.3. Recall that we aim to establish the following general expression for $k\geq 1$:

$$\mathrm{MSE}(\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y}))=$$
$$\frac{1}{m}\binom{n-m}{n-1}\left(((\mathbf{x}^\top\mathbf{y})^2+\|\mathbf{x}\|^2\|\mathbf{y}\|^2)+\sum_{r=1}^{k-1}\frac{(-1)^r2^r}{n^r}(2(\mathbf{x}^\top\mathbf{y})^2+\|\mathbf{x}\|^2\|\mathbf{y}\|^2)+\frac{(-1)^k2^k}{n^{k-1}}\sum_{i=1}^{n}x_i^2y_i^2\right).$$

We proceed by induction. The case $k=1$ is verified by Proposition 8.2. For the inductive step, suppose the result holds for some $k\in\mathbb{N}$. Then observe by Proposition 8.3 and the induction hypothesis, we have

$$\mathrm{MSE}(\widehat{K}_m^{(k+1)}(\mathbf{x},\mathbf{y})) = \mathbb{E}\left[\mathrm{MSE}\left(\widehat{K}_m^{(k-1)}(\mathbf{SD}_1\mathbf{x},\mathbf{SD}_1\mathbf{y})|\mathbf{D}_1\right)\right]$$
$$= \frac{1}{m}\binom{n-m}{n-1}\left(((\mathbf{x}^\top\mathbf{y})^2+\|\mathbf{x}\|^2\|\mathbf{y}\|^2)+\sum_{r=1}^{k-1}\frac{(-1)^r2^r}{n^r}(2(\mathbf{x}^\top\mathbf{y})^2+\|\mathbf{x}\|^2\|\mathbf{y}\|^2)\right.$$
$$\left.+\frac{(-1)^k2^k}{n^{k-1}}\sum_{i=1}^{n}\mathbb{E}\left[(\mathbf{SD}_1\mathbf{x})_i^2(\mathbf{SD}_1\mathbf{y})_i^2\right]\right),$$

where we have used that $\mathbf{SD}_1$ is almost surely orthogonal, and therefore $\|\mathbf{SD}_1\mathbf{x}\|^2=\|\mathbf{x}\|^2$ almost surely, $\|\mathbf{SD}_1\mathbf{y}\|^2=\|\mathbf{y}\|^2$ almost surely, and $\langle\mathbf{SD_1x},\mathbf{SD}_1\mathbf{y}\rangle=\langle\mathbf{x},\mathbf{y}\rangle$ almost surely. Applying Lemma 8.4 to the remaining expectation and collecting terms yields the required expression for $\mathrm{MSE}(\widehat{K}_m^{(k+1)}(\mathbf{x},\mathbf{y}))$, and the proof is complete. □

8.4 Proof of Lemma 3.5

Proof. Consider the last block $\mathbf{H}$ that is sub-sampled. Notice that if rows $\mathbf{r}^1$ and $\mathbf{r}^2$ of $\mathbf{H}$ of indices i and $\frac{n}{2}+i$ are chosen then from the recursive definition of $\mathbf{H}$ we conclude that $(\mathbf{r}^2)^\top\mathbf{x}=(\mathbf{r}_1^1)^\top\mathbf{x}-(\mathbf{r}_2^1)^\top\mathbf{x}$, where $\mathbf{r}_1^1,\mathbf{r}_2^1$ stand for the first and second half of $\mathbf{r}^1$ respectively. Thus computations of $(\mathbf{r}^1)^\top\mathbf{x}$ can be reused to compute both $(\mathbf{r}^1)^\top\mathbf{x}$ and $(\mathbf{r}^2)^\top\mathbf{x}$ in time $n+O(1)$ instead of $2n$. If we denote by r the expected number of pairs of rows $(i,\frac{n}{2}+i)$ that are chosen by the random sampling mechanism, then we see that by applying the trick above for all the r pairs, we obtain time complexity $O((k-1)n\log(n)+n(m-2r)+nr+r)$, where: $O((k-1)n\log(n))$ is the time required to compute first $(k-1)$ **HD** blocks (with the use of Walsh-Hadamard Transform), $O(n(m-2r))$ stands for time complexity of the brute force computations for these rows that were not coupled in the last block and $O(nr+r)$ comes from the above trick applied to all r aforementioned pairs of

rows. Thus, to obtain the first term in the min-expression on time complexity from the statement of the lemma, it remains to show that

$$\mathbb{E}[r] = \frac{(m-1)m}{2(n-1)}. \tag{28}$$

But this is straightforward. Note that the number of the m-subsets of the set of all n rows that contain some fixed rows of indices i_1, i_2 ($i_1 \neq i_2$) is $\binom{n-2}{m-2}$. Thus for any fixed pair of rows of indices i and $\frac{n}{2}+i$ the probability that these two rows will be selected is exactly $p_{succ} = \frac{\binom{n-2}{m-2}}{\binom{n}{m}} = \frac{(m-1)m}{(n-1)n}$. Equation 28 comes from the fact that clearly: $\mathbb{E}[r] = \frac{n}{2}p_{succ}$. Thus we obtain the first term in the min-expression from the statement of the lemma. The other one comes from the fact that one can always do all the computations by calculating k times Walsh-Hadamard transformation. That completes the proof.

□

8.5 Proof of Theorem 3.6

The proof of Theorem 3.6 follows a very similar structure to that of Theorem 3.3; we proceed by induction, and may use the results of Proposition 8.3 to set up a recursion. We first show unbiasedness of the estimator (Proposition 8.5), and then treat the base case of the inductive argument (Proposition 8.6). We prove slightly more general statements than needed for Theorem 3.6, as this will allow us to explore the fully complex case in §8.7.

Proposition 8.5. *The estimator $K_m^{\mathcal{H},(k)}(\mathbf{x},\mathbf{y})$ is unbiased for all $k, n \in \mathbb{N}$, $m \leq n$, and $\mathbf{x},\mathbf{y} \in \mathbb{C}^n$ with $\langle \overline{\mathbf{x}}, \mathbf{y}\rangle \in \mathbb{R}$; in particular, for all $\mathbf{x},\mathbf{y} \in \mathbb{R}$.*

Proof. Following a similar argument to the proof of Proposition 8.1, note that it is sufficient to prove the claim for $k = 1$, since each $\mathbf{SD}$ block is unitary, and hence preserves the Hermitian product $\langle \overline{\mathbf{x}}, \mathbf{y}\rangle$.

Next, note that the estimator can be written as a sum of identically distributed terms:

$$\widehat{K}_m^{\mathcal{H},(1)}(\mathbf{x},\mathbf{y}) = \frac{n}{m}\sum_{i=1}^{m} \mathrm{Re}\left((\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_i} \times (\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_i}\right) .$$

The terms are identically distributed since the index variables J_i are marginally identically distributed, and the rows of $\mathbf{SD}_1$ are marginally identically distributed (the elements of a row are iid $\mathrm{Unif}(S^1)/\sqrt{n}$). Now note

$$\begin{aligned}\mathbb{E}\left[\mathrm{Re}\left((\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_i} \times (\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_i}\right)\right] &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} y_i d_i \times \sum_{i=1}^{n} \overline{x}_i \overline{d}_i\right] \\ = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} \overline{x}_i y_i d_i \overline{d}_i\right] + \mathbb{E}\left[\sum_{i\neq j} \overline{x}_i y_j \overline{d}_i d_j\right] &= \frac{1}{n}\langle \overline{\mathbf{x}}, \mathbf{y}\rangle ,\end{aligned}$$

where $d_i = \mathbf{D}_{ii} \overset{iid}{\sim} \mathrm{Unif}(S^1)$ for $i = 1,\ldots,n$. This immediately yields $\mathbb{E}\left[\widehat{K}_m^{\mathcal{H},(1)}(\mathbf{x},\mathbf{y})\right] = \langle \overline{\mathbf{x}}, \mathbf{y}\rangle$, as required. □

We now derive the base case for our inductive proof, again proving a slightly more general statement then necessary for Theorem 3.6.

Proposition 8.6. *Let $\mathbf{x},\mathbf{y} \in \mathbb{C}^n$ such that $\langle \overline{\mathbf{x}}, \mathbf{y}\rangle \in \mathbb{R}$. The MSE of the single complex $\mathbf{SD}$-block m-feature estimator $K_m^{\mathcal{H},(1)}(\mathbf{x},\mathbf{y})$ for $\langle \overline{\mathbf{x}}, \mathbf{y}\rangle$ is*

$$\mathrm{MSE}(\widehat{K}_m^{\mathcal{H},(1)}(\mathbf{x},\mathbf{y})) = \frac{1}{2m}\left(\frac{n-m}{n-1}\right)\left(\langle \overline{\mathbf{x}}, \mathbf{x}\rangle\langle \overline{\mathbf{y}}, \mathbf{y}\rangle + \langle \overline{\mathbf{x}}, \mathbf{y}\rangle^2 - \sum_{r=1}^{n}|x_r|^2|y_r|^2 - \sum_{r=1}^{n}\mathrm{Re}(\overline{x}_r^2 y_r^2)\right) .$$

Proof. The proof is very similar to that of Proposition 8.2. By the unbiasedness result of Proposition 8.5, the mean squared error of the estimator is simply the variance. We begin by conditioning on the random index vector $\mathbf{J}$ selected by the sub-sampling procedure.

$$\widehat{K}_m^{\mathcal{H},(1)}(\mathbf{x},\mathbf{y})) = \frac{1}{M}\mathrm{Re}\left(\langle\sqrt{n}(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{\mathbf{J}}, \sqrt{n}(\mathbf{S}\mathbf{D}\mathbf{y})_{\mathbf{J}}\rangle\right) ,$$

where again $\mathbf{J}$ is a set of uniform iid indices from $1,\ldots,n$, and the bar over D represents complex conjugation. Since the estimator is again unbiased, its MSE is equal to its variance. First conditioning on the index set $\mathbf{J}$, as for Proposition 8.6, we obtain

$$\begin{aligned}&\mathrm{Var}\left(\widehat{K}_m^{\mathcal{H},(1)}(x,y)\right)\\ =&\frac{n^2}{m^2}\left(\mathbb{E}\left[\mathrm{Var}\left(\mathrm{Re}\left(\sum_{p=1}^m(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right)\right]+\mathrm{Var}\left(\mathbb{E}\left[\mathrm{Re}\left(\sum_{p=1}^m(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right]\right)\right).\end{aligned}$$

Again, the second term vanishes as the conditional expectation is constant as a function of $\mathbf{J}$, by unitarity of $\mathbf{SD}$. Turning attention to the conditional variance expression in the first term, we note

$$\begin{aligned}&\mathrm{Var}\left(\mathrm{Re}\left(\sum_{p=1}^m(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right) =\\ &\sum_{p,p'=1}^m\sum_{i,j,k,l=1}^n s_{J_pi}s_{J_pj}s_{J_{p'}k}s_{J_{p'}l}\mathrm{Cov}\left(\mathrm{Re}(\overline{d}_i\overline{x}_id_jy_j),\mathrm{Re}(\overline{d}_k\overline{x}_kd_ly_l)\right) .\end{aligned}$$

Now note that the covariance term is non-zero iff i,j are distinct, and $\{i,j\}=\{k,l\}$. We therefore obtain

$$\begin{aligned}&\mathrm{Var}\left(\mathrm{Re}\left(\sum_{p=1}^m(\mathbf{S}\overline{\mathbf{D}}\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right)\\ =&\sum_{p,p'=1}^m\sum_{i\neq j}^n s_{J_pi}s_{J_pj}s_{J_{p'}i}s_{J_{p'}j}\left(\mathrm{Cov}\left(\mathrm{Re}(\overline{d}_i\overline{x}_id_jy_j),\mathrm{Re}(\overline{d}_i\overline{x}_id_jy_j)\right)+\mathrm{Cov}\left(\mathrm{Re}(\overline{d}_i\overline{x}_id_jy_j),\mathrm{Re}(\overline{d}_j\overline{x}_jd_iy_i)\right)\right)\end{aligned}$$

First consider the term $\mathrm{Cov}\left(\mathrm{Re}(\overline{d}_i\overline{x}_id_jy_j),\mathrm{Re}(\overline{d}_i\overline{x}_id_jy_j)\right)$. The random variable $\overline{d}_i\overline{x}_id_jy_j$ is distributed uniformly on the circle in the complex plane centered at the origin with radius $|\overline{x}_iy_j|$. Therefore the variance of its real part is

$$\mathrm{Cov}\left(\mathrm{Re}(\overline{d}_i\overline{x}_id_jy_j),\mathrm{Re}(\overline{d}_i\overline{x}_id_jy_j)\right) = \frac{1}{2}|\overline{x}_iy_j|^2 = \frac{1}{2}x_i\overline{x}_iy_j\overline{y}_j .$$

For the second covariance term, we perform an explicit calculation. Let $Z=e^{i\theta}=\overline{d}_id_j$. Then we have

$$\begin{aligned}&\mathrm{Cov}\left(\mathrm{Re}(\overline{d}_i\overline{x}_id_jy_j),\mathrm{Re}(\overline{d}_j\overline{x}_jd_iy_i)\right) = \mathrm{Cov}\left(\mathrm{Re}(Z\overline{x}_iy_j),\mathrm{Re}(\overline{Z}\overline{x}_jy_i)\right)\\ &= \mathrm{Cov}\left(\cos(\theta)\mathrm{Re}(\overline{x}_iy_j)-\sin(\theta)\mathrm{Im}(\overline{x}_iy_j),\cos(\theta)\mathrm{Re}(\overline{x}_jy_i)+\sin(\theta)\mathrm{Im}(\overline{x}_jy_i)\right)\\ &= \frac{1}{2}\left(\mathrm{Re}(\overline{x}_iy_j)\mathrm{Re}(\overline{x}_jy_i)-\mathrm{Im}(\overline{x}_iy_j)\mathrm{Im}(\overline{x}_jy_i)\right) ,\end{aligned}$$

with the final equality following since the angle θ is uniformly distributed on $[0,2\pi]$, and standard trigonometric integral identities. We recognize the bracketed terms in the final line as the real part of the product $\overline{x}_i\overline{x}_jy_iy_j$. Substituting these into the expression for the conditional variance obtained above, we have

$$\mathrm{Var}\left(\mathrm{Re}\left(\sum_{p=1}^m(\mathbf{S}\overline{\mathbf{D}}\mathbf{x})_{J_p}(\mathbf{S}\mathbf{D}\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right) = \sum_{p,p'=1}^m\sum_{i\neq j}^n s_{J_pi}s_{J_pj}s_{J_{p'}i}s_{J_{p'}j}\frac{1}{2}\left(x_i\overline{x}_iy_j\overline{y}_j+\mathrm{Re}(\overline{x}_i\overline{x}_jy_iy_j)\right).$$

Now taking the expectation over the index variables $\mathbf{J}$, we note that as in the proof of Proposition 8.2, the expectation of the term $s_{J_pi}s_{J_pj}s_{J_{p'}i}s_{J_{p'}j}$ is $1/n^2$ when $p=p'$, and $1/(n^2(n-1))$ otherwise. Therefore we obtain

$$\mathrm{Var}\left(\widehat{K}_m^{\mathcal{H},(1)}(\mathbf{x},\mathbf{y})\right) = \frac{n^2}{m^2}\left(\left(\frac{m}{n^2}+\frac{m(m-1)}{n^2(n-1)}\right)\frac{1}{2}\sum_{i\neq j}^n\left(x_i\overline{x}_iy_j\overline{y}_j+\mathrm{Re}(\overline{x}_i\overline{x}_jy_iy_j)\right)\right)$$

$$= \frac{1}{2m}\left(\frac{n-m}{n-1}\right)\left(\sum_{i\neq j}^{n}\left(x_i\overline{x}_i y_j\overline{y}_j + \mathrm{Re}(\overline{x}_i\overline{x}_j y_i y_j)\right)\right)$$
$$= \frac{1}{2m}\left(\frac{n-m}{n-1}\right)\left(\sum_{i,j=1}^{n}\left(x_i\overline{x}_i y_j\overline{y}_j + \mathrm{Re}(\overline{x}_i\overline{x}_j y_i y_j)\right) - \sum_{i=1}^{n}(x_i\overline{x}_i y_i\overline{y}_i + \mathrm{Re}(\overline{x}_i\overline{x}_i y_i y_i))\right)$$
$$= \frac{1}{2m}\left(\frac{n-m}{n-1}\right)\left(\langle\overline{\mathbf{x}},\mathbf{x}\rangle\langle\overline{\mathbf{y}},\mathbf{y}\rangle + \langle\overline{\mathbf{x}},\mathbf{y}\rangle^2 - \sum_{i=1}^{n}(x_i\overline{x}_i y_i\overline{y}_i + \mathrm{Re}(\overline{x}_i\overline{x}_i y_i y_i))\right),$$

where in the final equality we have used the assumption that $\langle\overline{\mathbf{x}},\mathbf{y}\rangle \in \mathbb{R}$. □

We are now in a position to prove Theorem 3.6 by induction, using Proposition 8.6 as a base case, and Proposition 8.3 for the inductive step.

Proof of Theorem 3.6. Recall that we aim to establish the following general expression for $k \geq 1$:

$$\mathrm{MSE}(\widehat{K}_m^{\mathcal{H},(k)}(\mathbf{x},\mathbf{y})) = \frac{1}{2m}\left(\frac{n-m}{n-1}\right)\Bigg(((\mathbf{x}^\top\mathbf{y})^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \sum_{r=1}^{k-1}\frac{(-1)^r 2^r}{n^r}(2(\mathbf{x}^\top\mathbf{y})^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \frac{(-1)^k 2^k}{n^{k-1}}\sum_{i=1}^{n}x_i^2 y_i^2\Bigg).$$

We proceed by induction. The case $k = 1$ is verified by Proposition 8.6, and by noting that in the expression obtained in Proposition 8.6, we have

$$\sum_{i=1}^{n} x_i\overline{x}_i y_i\overline{y}_i = \mathrm{Re}(\overline{x}_i\overline{x}_i y_i y_i) = \sum_{i=1}^{n} x_i^2 y_i^2\,.$$

For the inductive step, suppose the result holds for some $k \in \mathbb{N}$. Then observe by Proposition 8.3 and the induction hypothesis, we have, for $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$:

$$\mathrm{MSE}(\widehat{K}_m^{\mathcal{H},(k+1)}(\mathbf{x},\mathbf{y})) = \mathbb{E}\left[\mathrm{MSE}\left(\widehat{K}_m^{(k-1)}(\mathbf{SD}_1\mathbf{x},\mathbf{SD}_1\mathbf{y})|\mathbf{D}_1\right)\right]$$
$$= \frac{1}{2m}\left(\frac{n-m}{n-1}\right)\Bigg(((\mathbf{x}^\top\mathbf{y})^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \sum_{r=1}^{k-1}\frac{(-1)^r 2^r}{n^r}(2(\mathbf{x}^\top\mathbf{y})^2 + \|\mathbf{x}\|^2\|\mathbf{y}\|^2) + \frac{(-1)^k 2^k}{n^{k-1}}\sum_{i=1}^{n}\mathbb{E}\left[(\mathbf{SD}_1\mathbf{x})_i^2(\mathbf{SD}_1\mathbf{y})_i^2\right]\Bigg),$$

where we have used that $\mathbf{SD}_1$ is almost surely orthogonal, and therefore $\|\mathbf{SD}_1\mathbf{x}\|^2 = \|\mathbf{x}\|^2$ almost surely, $\|\mathbf{SD}_1\mathbf{y}\|^2 = \|\mathbf{y}\|^2$ almost surely, and $\langle\mathbf{SD_1x},\mathbf{SD}_1\mathbf{y}\rangle = \langle\mathbf{x},\mathbf{y}\rangle$ almost surely. Applying Lemma 8.4 to the remaining expectation and collecting terms yields the required expression for $\mathrm{MSE}(\widehat{K}_m^{\mathcal{H},(k+1)}(\mathbf{x},\mathbf{y}))$, and the proof is complete.

□

8.6 Proof of Corollary 3.7

The proof follows simply by following the inductive strategy of the proof of Theorem 3.6, replacing the base case in Proposition 8.6 with the following.

Proposition 8.7. *Let $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$. The MSE of the single hybrid $\mathbf{SD}$-block m-feature estimator $K_m^{\mathcal{H},(1)}(\mathbf{x},\mathbf{y})$ using a diagonal matrix with entries $\mathrm{Unif}(\{1,-1,i,-i\})$, rather than $\mathrm{Unif}(S^1)$ for $\langle\mathbf{x},\mathbf{y}\rangle$ is*

$$\mathrm{MSE}(\widehat{K}_m^{\mathcal{H},(1)}(\mathbf{x},\mathbf{y})) = \frac{1}{2m}\left(\langle\overline{\mathbf{x}},\mathbf{x}\rangle\langle\overline{\mathbf{y}},\mathbf{y}\rangle + \langle\overline{\mathbf{x}},\mathbf{y}\rangle^2 - 2\sum_{r=1}^{n}x_r^2 y_r^2\right).$$

Proof. The proof of this proposition proceeds exactly as for Proposition 8.6; by following the same chain of reasoning, conditioning on the index set $\mathbf{J}$ of the sub-sampled rows, we arrive at

$$\mathrm{Var}\left(\mathrm{Re}\left(\sum_{p=1}^{m}(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right) = \sum_{p,p'=1}^{m}\sum_{i,j,k,l=1}^{n} s_{J_p i}s_{J_p j}s_{J_{p'}k}s_{J_{p'}l}\mathrm{Cov}\left(\mathrm{Re}(\overline{d}_i\overline{x}_i d_j y_j),\mathrm{Re}(\overline{d}_k\overline{x}_k d_l y_l)\right) .$$

Since we are dealing strictly with the case $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, we may simplify this further to obtain

$$\mathrm{Var}\left(\mathrm{Re}\left(\sum_{p=1}^{m}(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right) = \sum_{p,p'=1}^{m}\sum_{i,j,k,l=1}^{n} s_{J_p i}s_{J_p j}s_{J_{p'}k}s_{J_{p'}l}x_i x_k y_i y_l\mathrm{Cov}\left(\mathrm{Re}(\overline{d}_i d_j),\mathrm{Re}(\overline{d}_k d_l)\right) .$$

By calculating directly with the $d_i,d_j,d_k,d_l\sim\mathrm{Unif}(\{1,-1,i,-i\})$, we obtain

$$\mathrm{Var}\left(\mathrm{Re}\left(\sum_{p=1}^{m}(\mathbf{S}\overline{\mathbf{D}}_1\overline{\mathbf{x}})_{J_p}(\mathbf{S}\mathbf{D}_1\mathbf{y})_{J_p}\right)\middle|\mathbf{J}\right) = \frac{1}{2}\sum_{p,p'=1}^{m}\sum_{i\neq j}^{n} s_{J_p i}s_{J_p j}s_{J_{p'}k}s_{J_{p'}l}(x_i^2y_j^2+x_ix_jy_iy_j) ,$$

exactly as in Proposition 8.6; following the rest of the argument of Proposition 8.6 yields the result. □

The proof of the corollary now follows by applying the steps of the proof of Theorem 3.6.

8.7 Exploring Dimensionality Reduction with Fully-complex Random Matrices

In this section, we briefly explore the possibility of using $\mathbf{SD}$-product matrices in which all the random diagonal matrices are complex-valued. Following on from the ROMs introduced in Definition 2.1, we define the $\mathbf{S}$-*Uniform* random matrix with $k\in\mathbb{N}$ blocks to be given by

$$\mathbf{M}_{\mathbf{S}\mathcal{U}}^{(k)} = \prod_{i=1}^{k}\mathbf{S}\mathbf{D}_i^{(\mathcal{U})} ,$$

where $(\mathbf{D}_i^{(\mathcal{U})})_{i=1}^{k}$ are iid diagonal matrices with iid $\mathrm{Unif}(S^1)$ random variables on the diagonals, and S^1 is the unit circle of $\mathbb{C}$.

As alluded to in §3, we will see that introducing this increased number of complex parameters does not lead to significant increases in statistical performance relative to the estimator $\widehat{K}_m^{\mathcal{H},(k)}$ for dimensionality reduction.

We consider the estimator $\widehat{K}_m^{\mathcal{U},(k)}$ below, based on the sub-sampled $\mathbf{SD}$-product matrix $\mathbf{M}_{\mathbf{S}\mathcal{U}}^{(k),\mathrm{sub}}$:

$$\widehat{K}_m^{\mathcal{U},(k)}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\mathrm{Re}\left[\left(\overline{\mathbf{M}_{\mathbf{S}\mathcal{U}}^{(k),\mathrm{sub}}\mathbf{x}}\right)^{\top}\left(\mathbf{M}_{\mathbf{S}\mathcal{U}}^{(k),\mathrm{sub}}\mathbf{y}\right)\right] ,$$

and show that it does not yield a significant improvement over the estimator $\widehat{K}_m^{\mathcal{H},(k)}$ of Theorem 3.6:

Theorem 8.8. *For $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, the estimator $\widehat{K}_m^{\mathcal{U},(k)}(\mathbf{x},\mathbf{y})$, applying random sub-sampling strategy without replacement is unbiased and satisfies:*

$$\mathrm{MSE}(\widehat{K}_m^{\mathcal{U},(k)}(\mathbf{x},\mathbf{y})) = \frac{1}{2m}\left(\frac{n-m}{n-1}\right)\left(\left((\mathbf{x}^{\top}\mathbf{y})^2+\|\mathbf{x}\|^2\|\mathbf{y}\|^2\right)+\sum_{r=1}^{k-1}\frac{(-1)^r}{n^r}(3(\mathbf{x}^{\top}\mathbf{y})^2+\|\mathbf{x}\|^2\|\mathbf{y}\|^2)+\frac{(-1)^k 2}{n^{k-1}}\sum_{i=1}^{n}x_i^2y_i^2\right).$$

The structure of the proof of Theorem 8.8 is broadly the same as that of Theorem 3.3. We begin by remarking that the proof that the estimator is unbiased is exactly the same as that of Proposition 8.5. We then note that in the case of $k = 1$ block, the estimators $\widehat{K}_m^{\mathcal{H},(1)}$ and $\widehat{K}_m^{\mathcal{U},(1)}$, coincide so Proposition 8.6 establishes the MSE of the estimator $\widehat{K}_m^{\mathcal{U},(k)}$ in the base case $k = 1$. We then obtain a recursion formula for the MSE (Proposition 8.9), and finally prove the theorem by induction.

Proposition 8.9. *Let $k \geq 2$, $n \in \mathbb{N}$, $m \leq n$, and $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ such that $\langle \overline{\mathbf{x}}, \mathbf{y} \rangle \in \mathbb{R}$; in particular, this includes $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Then we have the following recursion for the MSE of $\widehat{K}_M^{\mathcal{U},(k)}(\mathbf{x}, \mathbf{y})$:*

$$\mathrm{MSE}(\widehat{K}_m^{\mathcal{U},(k)}(\mathbf{x}, \mathbf{y})) = \mathbb{E}\left[\mathrm{MSE}(\widehat{K}_m^{\mathcal{U},(k-1)}(\mathbf{SD}_1\mathbf{x}, \mathbf{SD}_1\mathbf{y})\big|\mathbf{D}_1)\right]$$

Proof. The proof is exactly analogous to that of Proposition 8.3, and is therefore omitted. □

Before we complete the proof by induction, we will need the following auxiliary result, to deal with the expectations that arise during the recursion due to the terms in the MSE expression of Proposition 8.6.

Lemma 8.10. *Under the assumptions of Theorem 8.8, we have the following expectations:*

$$\mathbb{E}\left[|(\mathbf{SDx})_r|^2|(\mathbf{SDy})_r|^2\right] = \frac{1}{n^2}\left(\langle \overline{\mathbf{x}}, \mathbf{x}\rangle\langle \overline{\mathbf{y}}, \mathbf{y}\rangle + \langle \overline{\mathbf{x}}, \mathbf{y}\rangle^2 - \sum_{i=1}^n |x_i|^2|y_i|^2\right)$$

$$\mathbb{E}\left[\mathrm{Re}((\mathbf{S}\overline{\mathbf{D}}\overline{\mathbf{x}})_r^2(\mathbf{SDy})_r^2)\right] = \frac{1}{n^2}\left(2\langle \overline{\mathbf{x}}, \mathbf{y}\rangle^2 - \sum_{i=1}^n \mathrm{Re}(\overline{x}_i^2 y_i^2)\right)$$

Proof. For the first claim, we note that

$$\begin{aligned}
\mathbb{E}\left[|(\mathbf{SDx})_r|^2|(\mathbf{SDy})_r|^2\right] &= \sum_{i,j,k,l}^n s_{ri}s_{rj}s_{rk}s_{rl}\overline{x}_i x_j \overline{y}_k y_l \mathbb{E}\left[\overline{d}_i d_j \overline{d}_k d_l\right] \\
&= \frac{1}{n^2}\left(\sum_{i\neq j}\overline{x}_i x_i \overline{y}_j y_j + \sum_{i\neq j}\overline{x}_i x_j \overline{y}_j y_i + \sum_{i=1}^n \overline{x}_i x_i \overline{y}_i y_i\right) \\
&= \frac{1}{n^2}\left(\sum_{i,j=1}^n \overline{x}_i x_i \overline{y}_j y_j + \sum_{i,j=1}^n \overline{x}_i x_j \overline{y}_j y_i - \sum_{i=1}^n \overline{x}_i x_i \overline{y}_i y_i\right) \\
&= \frac{1}{n^2}\left(\langle \overline{\mathbf{x}}, \mathbf{x}\rangle\langle \overline{\mathbf{y}}, \mathbf{y}\rangle + \langle \overline{\mathbf{x}}, \mathbf{y}\rangle^2 - \sum_{i=1}^n |x_i|^2|y_i|^2\right),
\end{aligned}$$

as required, where in the final equality we have use the assumption that $\langle \overline{\mathbf{x}}, \mathbf{y}\rangle \in \mathbb{R}$. For the second claim, we observe that

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Re}((\overline{\mathbf{SDx}})_r^2(\mathbf{SDy})_r^2\right] &= \mathrm{Re}\left(\sum_{i,j,k,l}^n s_{ri}s_{rj}s_{rk}s_{rl}\overline{x}_i\overline{x}_j y_k y_l \mathbb{E}\left[\overline{d}_i\overline{d}_j d_k d_l\right]\right) \\
&= \mathrm{Re}\left(\frac{1}{n^2}\left(2\sum_{i\neq j}\overline{x}_i\overline{x_j} y_i y_j + \sum_{i=1}^n \overline{x}_i\overline{x}_i y_i y_i\right)\right) \\
&= \frac{1}{n^2}\left(2\langle \overline{\mathbf{x}}, \mathbf{y}\rangle^2 - \sum_{i=1}^n \mathrm{Re}\left(\overline{x}_i^2 y_i^2\right)\right),
\end{aligned}$$

where again we have used the assumption that $\langle \overline{\mathbf{x}}, \mathbf{y}\rangle \in \mathbb{R}$. □

Proof of Theorem 8.8. The proof now proceeds by induction. We in fact prove the stronger result that for any $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ for which $\langle \overline{\mathbf{x}}, \mathbf{y}\rangle \in \mathbb{R}$, we have

$$\mathrm{MSE}(\widehat{K}_m^{\mathcal{U},(k)}(\mathbf{x}, \mathbf{y})) = \frac{1}{2m}\left(\frac{n-m}{n-1}\right)\left(\left(\langle \overline{\mathbf{x}}, \mathbf{y}\rangle^2 + \langle \overline{\mathbf{x}}, \mathbf{x}\rangle\langle \overline{\mathbf{y}}, \mathbf{y}\rangle\right) + \sum_{r=1}^{k-1}\frac{(-1)^r}{n^r}(3\langle \overline{\mathbf{x}}, \mathbf{y}\rangle^2 + \langle \overline{\mathbf{x}}, \mathbf{x}\rangle\langle \overline{\mathbf{y}}, \mathbf{y}\rangle) +\right.$$

$$\frac{(-1)^k}{n^{k-1}}\left(\sum_{i=1}^{n}\left(|x_i|^2|y_i|^2+\mathrm{Re}\left(\overline{x}_i^2y_i^2\right)\right)\right)\Bigg).$$

from which Theorem 8.8 clearly follows. Proposition 8.6 yields the base case $k=1$ for this claim. For the recursive step, suppose that the result holds for some number $k\in\mathbb{N}$ of blocks. Recalling the recursion of Proposition 8.9, we then obtain

$$\mathrm{MSE}(\widehat{K}_m^{\mathcal{U},(k+1)}(\mathbf{x},\mathbf{y}))=\frac{1}{2m}\left(\frac{n-m}{n-1}\right)\Bigg(\left(\langle\overline{\mathbf{x}},\mathbf{y}\rangle^2+\langle\overline{\mathbf{x}},\mathbf{x}\rangle\langle\overline{\mathbf{y}},\mathbf{y}\rangle\right)+\sum_{r=1}^{k-1}\frac{(-1)^r}{n^r}(3\langle\overline{\mathbf{x}},\mathbf{y}\rangle^2+\langle\overline{\mathbf{x}},\mathbf{x}\rangle\langle\overline{\mathbf{y}},\mathbf{y}\rangle)+$$
$$\frac{(-1)^k}{n^{k-1}}\left(\sum_{i=1}^{n}\left(\mathbb{E}\left[|\mathbf{SD}_1\mathbf{x}|_i^2|\mathbf{SD}_1\mathbf{y}|_i^2\right]+\mathbb{E}\left[\mathrm{Re}\left((\overline{\mathbf{SD}_1\mathbf{x}})_i^2(\mathbf{SD}_1\mathbf{y})_i^2\right)\right]\right)\right)\Bigg),$$

where we have used the fact that $\mathbf{SD}_1$ is a unitary isometry almost surely, and thus preserves Hermitian products. Applying Lemma 8.10 to the remaining expectations and collecting terms proves the inductive step, which concludes the proof of the theorem. □

8.8 Proof of Theorem 3.8

Proof. The proof of this result is reasonably straightforward with the proofs of Theorems 3.3 and 3.6 in hand; we simply recognize where in these proofs the assumption of the sampling strategy without replacement was used. We deal first with Theorem 3.3, which deals with the MSE associated with $\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y})$. The only place in which the assumption of the sub-sampling strategy without replacement is used is mid-way through the proof of Proposition 8.2, which quantifies $\mathrm{MSE}(\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y}))$. Picking up the proof at the point the sub-sampling strategy is used, we have

$$\mathrm{MSE}(\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y}))=\frac{n^2}{m^2}\sum_{p,p'=1}^{m}\sum_{i\neq j}^{n}\left(x_i^2y_j^2+x_ix_jy_iy_j\right)\mathbb{E}\left[s_{J_pi}s_{J_pj}s_{J_{p'}i}s_{J_{p'}j}\right].$$

Now instead using sub-sampling strategy with replacement, note that each pair of sub-sampled indices J_p and $J_{p'}$ are independent. Recalling that the columns of $\mathbf{S}$ are orthogonal, we obtain for distinct p and p' that

$$\mathbb{E}\left[s_{J_pi}s_{J_pj}s_{J_{p'}i}s_{J_{p'}j}\right]=\mathbb{E}\left[s_{J_pi}s_{J_pj}\right]\mathbb{E}\left[s_{J_{p'}i}s_{J_{p'}j}\right]=0\,.$$

Again, for $p=p'$, we have $\mathbb{E}\left[s_{J_pi}s_{J_pj}s_{J_{p'}i}s_{J_{p'}j}\right]=1/n^2$. Substituting the values of these expectations back into the expression for the MSE of $\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y})$ then yields

$$\begin{aligned}\mathrm{MSE}(\widehat{K}_m^{(1)}(\mathbf{x},\mathbf{y}))=&\frac{n^2}{m^2}\sum_{i\neq j}^{n}\left(x_i^2y_j^2+x_ix_jy_iy_j\right)\left(m\times\frac{1}{n^2}\right)\\=&\frac{1}{m}\left(1-\frac{m-1}{n-1}\right)\sum_{i\neq j}^{n}\left(x_i^2y_j^2+x_ix_jy_iy_j\right)\\=&\frac{1}{m}\left(\langle\mathbf{x},\mathbf{y}\rangle^2+\|\mathbf{x}\|^2\|\mathbf{y}\|^2-2\sum_{i=1}^{n}x_i^2y_i^2\right)\end{aligned}$$

as required.

For the estimator $\widehat{K}_m^{\mathcal{H},(k)}(\mathbf{x},\mathbf{y})$, the result also immediately follows with the above calculation, as the only point in the proof of the MSE expressions for these estimators that is influenced by the sub-sampling strategy is in the calculation of the quantities $\mathbb{E}\left[s_{J_pi}s_{J_pj}s_{J_{p'}i}s_{J_{p'}j}\right]$; therefore, exactly the same multiplicative factor is incurred for MSE as for $\widehat{K}_m^{(k)}(\mathbf{x},\mathbf{y})$.

□

9 Proofs of results in §4

9.1 Proof of Lemma 4.2

Proof. Follows immediately from the proof of Theorem 4.4 (see: the proof below). □

9.2 Proof of Theorem 4.3

Recall that the angular kernel estimator based on $\mathbf{G}_{\mathrm{ort}}$ is given by

$$\widehat{K}_m^{\mathrm{ang,ort}}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\mathrm{sign}(\mathbf{G}_{\mathrm{ort}}\mathbf{x})^\top \mathrm{sign}(\mathbf{G}_{\mathrm{ort}}\mathbf{y})$$

where the function sign acts on vectors element-wise. In what follows, we write $\mathbf{G}_{\mathrm{ort}}^i$ for the ith row of $\mathbf{G}_{\mathrm{ort}}$, and $\mathbf{G}_i$ for the ith row of $\mathbf{G}$.

Since each $\mathbf{G}_{\mathrm{ort}}^i$ has the same marginal distribution as R_m in the unstructured Gaussian case covered by Theorem 4.4, unbiasedness of $\widehat{K}^{\mathrm{ang,ort}}(x,y)$ follows immediately from this result, and so we obtain:

Lemma 9.1. *$\widehat{K}_m^{\mathrm{ang,ort}}(\mathbf{x},\mathbf{y})$ is an unbiased estimator of $K^{\mathrm{ang}}(\mathbf{x},\mathbf{y})$.*

We now turn our attention to the variance of $\widehat{K}_m^{\mathrm{ang,ort}}(\mathbf{x},\mathbf{y})$.

Theorem 9.2. *The variance of the estimator $\widehat{K}_m^{\mathrm{ang,ort}}(x,y)$ is strictly smaller than the variance of $\widehat{K}_m^{\mathrm{ang,\ base}}(\mathbf{x},\mathbf{y})$*

Proof. Denote by θ the angle between $\mathbf{x}$ and $\mathbf{y}$, and for notational ease, let $S_i = \mathrm{sign}\left(\langle \mathbf{G}^i, \mathbf{x}\rangle\right)\mathrm{sign}\left(\langle \mathbf{G}^i, \mathbf{y}\rangle\right)$, and $S_i^{\mathrm{ort}} = \mathrm{sign}\left(\langle \mathbf{G}_{\mathrm{ort}}^i, \mathbf{x}\rangle\right)\mathrm{sign}\left(\langle \mathbf{G}_{\mathrm{ort}}^i, \mathbf{y}\rangle\right)$. Now observe that as $\widehat{K}_m^{\mathrm{ang,ort}}(\mathbf{x},\mathbf{y})$ is unbiased, we have

$$\begin{aligned}
&\mathrm{Var}\left(\widehat{K}_m^{\mathrm{ang,ort}}(\mathbf{x},\mathbf{y})\right)\\
&= \mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^m S_i^{\mathrm{ort}}\right)\\
&= \frac{1}{m^2}\left(\sum_{i=1}^m \mathrm{Var}\left(S_i^{\mathrm{ort}}\right) + \sum_{i\neq i'}^m \mathrm{Cov}\left(S_i^{\mathrm{ort}}, S_{i'}^{\mathrm{ort}}\right)\right).
\end{aligned}$$

By a similar argument, we have

$$\mathrm{Var}\left(\widehat{K}_m^{\mathrm{base}}(\mathbf{x},\mathbf{y})\right) = \frac{1}{m^2}\left(\sum_{i=1}^m \mathrm{Var}\left(S_i\right) + \sum_{i\neq i'}^m \mathrm{Cov}\left(S_i, S_{i'}\right)\right). \tag{29}$$

Note that the covariance terms in (29) evaluate to 0, by independence of S_i and $S_{i'}$ for $i \neq i'$ (which is inherited from the independence of $\mathbf{G}^i$ and $\mathbf{G}^{i'}$). Also observe that since $\mathbf{G}^i \overset{d}{=} \mathbf{G}_{\mathrm{ort}}^i$, we have

$$\mathrm{Var}\left(S_i^{\mathrm{ort}}\right) = \mathrm{Var}\left(S_i\right).$$

Therefore, demonstrating the theorem is equivalent to showing, for $i \neq i'$, that

$$\mathrm{Cov}\left(S_i^{\mathrm{ort}}, S_{i'}^{\mathrm{ort}}\right) < 0\,,$$

which is itself equivalent to showing

$$\mathbb{E}\left[S_i^{\mathrm{ort}} S_{i'}^{\mathrm{ort}}\right] < \mathbb{E}\left[S_i^{\mathrm{ort}}\right]\mathbb{E}\left[S_{i'}^{\mathrm{ort}}\right]. \tag{30}$$

Note that the variables $(S_i^{\mathrm{ort}})_{i=1}^m$ take values in $\{\pm 1\}$. Denoting $\mathcal{A}_i = \{S_i^{\mathrm{ort}} = -1\}$ for $i = 1,\ldots,m$, we can rewrite (30) as

$$\mathbb{P}\left[\mathcal{A}_i^c \cap \mathcal{A}_{i'}^c\right] + \mathbb{P}\left[\mathcal{A}_i \cap \mathcal{A}_{i'}\right] - \mathbb{P}\left[\mathcal{A}_i \cap \mathcal{A}_{i'}^c\right] - \mathbb{P}\left[\mathcal{A}_i^c \cap \mathcal{A}_{i'}\right] < \left(\frac{\pi - 2\theta}{\pi}\right)^2.$$

Note that the left-hand side is equal to

$$2(\mathbb{P}\left[\mathcal{A}_i^c \cap \mathcal{A}_{i'}^c\right] + \mathbb{P}\left[\mathcal{A}_i \cap \mathcal{A}_{i'}\right]) - 1\,.$$

Plugging in the bounds of Proposition 9.3, and using the fact that the pair of indicators $(\mathbb{1}_{\mathcal{A}_i}, \mathbb{1}_{\mathcal{A}_{i'}})$ is identically distributed for all pairs of distinct indices $i, i' \in \{1,\ldots,m\}$, thus yields the result. □

Proposition 9.3. *We then have the following inequalities:*

$$\mathbb{P}\left[\mathcal{A}_1 \cap \mathcal{A}_2\right] < \left(\frac{\theta}{\pi}\right)^2 \qquad \text{and} \qquad \mathbb{P}\left[\mathcal{A}_1^c \cap \mathcal{A}_2^c\right] < \left(1 - \frac{\theta}{\pi}\right)^2$$

Before providing the proof of this proposition, we describe some coordinate choices we will make in order to obtain the bounds in Proposition 9.3.

We pick an orthonormal basis for $\mathbb{R}^n$ so that the first two coordinates span the $\mathbf{x}$-$\mathbf{y}$ plane, and further so that $(\mathbf{G}^1_{\text{ort}})_2$, the coordinate of $\mathbf{G}^1_{\text{ort}}$ in the second dimension, is 0. We extend this to an orthonormal basis of $\mathbb{R}^n$ so that $(\mathbf{G}^1_{\text{ort}})_3 \geq 0$, and $(\mathbf{G}^1_{\text{ort}})_i = 0$ for $i \geq 4$. Thus, in this basis, we have coordinates

$$\mathbf{G}^1_{\text{ort}} = \left((\mathbf{G}^1_{\text{ort}})_1, 0, (\mathbf{G}^1_{\text{ort}})_3, 0, \ldots, 0\right),$$

with $(\mathbf{G}^1_{\text{ort}})_1 \sim \chi_2$ and $(\mathbf{G}^1_{\text{ort}})_3 \sim \chi_{N-2}$ (by elementary calculations with multivariate Gaussian distributions). Note that the angle, ϕ, that $\mathbf{G}^1_{\text{ort}}$ makes with the $\mathbf{x}$-$\mathbf{y}$ plane is then $\phi = \arctan((\mathbf{G}^1_{\text{ort}})_3/(\mathbf{G}^1_{\text{ort}})_1)$. Having fixed our coordinate system relative to the random variable $\mathbf{G}^1_{\text{ort}}$, the coordinates of $\mathbf{x}$ and $\mathbf{y}$ in this frame are now themselves random variables; we introduce the angle ψ to describe the angle between $\mathbf{x}$ and the positive first coordinate axis in this basis.

Now consider $\mathbf{G}^2_{\text{ort}}$. We are concerned with the direction of $((\mathbf{G}^2_{\text{ort}})_1, (\mathbf{G}^2_{\text{ort}})_2)$ in the $\mathbf{x}$-$\mathbf{y}$ plane. Conditional on $\mathbf{G}^1_{\text{ort}}$, the direction of the full vector $\mathbf{G}^2_{\text{ort}}$ is distributed uniformly on $S^{n-2}(\langle \mathbf{G}^1_{\text{ort}} \rangle^{\perp})$, the set of unit vectors orthogonal to $\mathbf{G}^1_{\text{ort}}$. Because of our particular choice of coordinates, we can therefore write

$$\mathbf{G}^2_{\text{ort}} = \left(r\sin(\phi), (\mathbf{G}^2_{\text{ort}})_2, r\cos(\phi), (\mathbf{G}^2_{\text{ort}})_4, (\mathbf{G}^2_{\text{ort}})_5, \ldots, (\mathbf{G}^2_{\text{ort}})_n\right),$$

where the $(N-1)$-dimensional vector $(r, (\mathbf{G}^2_{\text{ort}})_2, (\mathbf{G}^2_{\text{ort}})_4, (\mathbf{G}^2_{\text{ort}})_5, \ldots, (\mathbf{G}^2_{\text{ort}})_n)$ has an isotropic distribution.

So the direction of $((\mathbf{G}^2_{\text{ort}})_1, (\mathbf{G}^2_{\text{ort}})_2)$ in the $\mathbf{x}$-$\mathbf{y}$ plane follows an angular Gaussian distribution, with covariance matrix

$$\begin{pmatrix} \sin^2(\phi) & 0 \\ 0 & 1 \end{pmatrix}.$$

With these geometrical considerations in place, we are ready to give the proof of Proposition 9.3.

Proof of Proposition 9.3. Dealing with the first inequality, we decompose the event as

$$\begin{aligned}
\mathcal{A}_1 \cap \mathcal{A}_2 =& \{\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0\} \\
&\cup \{\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle > 0\} \\
&\cup \{\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle < 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0\} \\
&\cup \{\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle < 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle > 0\}.
\end{aligned}$$

As the law of $(\mathbf{G}^1_{\text{ort}}, \mathbf{G}^2_{\text{ort}})$ is the same as that of $(\mathbf{G}^2_{\text{ort}}, \mathbf{G}^1_{\text{ort}})$ and that of $(-\mathbf{G}^1_{\text{ort}}, \mathbf{G}^2_{\text{ort}})$, it follows that all four events in the above expression have the same probability. The statement of the theorem is therefore equivalent to demonstrating the following inequality:

$$\mathbb{P}\left[\langle \mathbf{G}^1_{\text{ort}}, x\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0\right] < \left(\frac{\theta}{2\pi}\right)^2.$$

We now proceed according to the coordinate choices described above. We first condition on the random angles ϕ and ψ to obtain

$$\begin{aligned}
&\mathbb{P}\left[\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0\right] \\
&= \int_0^{2\pi} \frac{\mathrm{d}\psi}{2\pi} \int_0^{\pi/2} f(\phi)\mathrm{d}\phi\, \mathbb{P}\left[\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0 | \psi, \phi\right] \\
&= \int_0^{2\pi} \frac{\mathrm{d}\psi}{2\pi} \int_0^{\pi/2} f(\phi)\mathrm{d}\phi\, \mathbb{1}_{\{0 \in [\psi - \pi/2, \psi - \pi/2 + \theta]\}} \mathbb{P}\left[\langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0 | \psi, \phi\right],
\end{aligned}$$

where f is the density of the random angle ϕ. The final equality above follows as $\mathbf{G}^1_{\text{ort}}$ and $\mathbf{G}^2_{\text{ort}}$ are independent conditional on ψ and ϕ, and since the event $\{\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0\}$ is exactly the event $\{0 \in [\psi - \pi/2, \psi - \pi/2 + \theta]\}$, by considering the geometry of the situation in the $\mathbf{x}$-$\mathbf{y}$ plane. We can remove the indicator function from the integrand by adjusting the limits of integration, obtaining

$$\begin{aligned}
&\mathbb{P}\left[\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0\right] \\
&= \int_{\pi/2-\theta}^{\pi/2} \frac{\mathrm{d}\psi}{2\pi} \int_0^{\pi/2} f(\phi)\mathrm{d}\phi\, \mathbb{P}\left[\langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0 | \psi, \phi\right] .
\end{aligned}$$

We now turn our attention to the conditional probability

$$\mathbb{P}\left[\langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0 | \psi, \phi\right] .$$

The event $\{\langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0\}$ is equivalent to the angle t of the projection of $\mathbf{G}^2_{\text{ort}}$ into the $\mathbf{x}$-$\mathbf{y}$ plane with the first coordinate axis lying in the interval $[\psi - \pi/2, \psi - \pi/2 + \theta]$. Recalling the distribution of the angle t from the geometric considerations described immediately before this proof, we obtain

$$\begin{aligned}
&\mathbb{P}\left[\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0\right] \\
&= \int_{\pi/2-\theta}^{\pi/2} \frac{\mathrm{d}\psi}{2\pi} \int_0^{\pi/2} f(\phi)\mathrm{d}\phi \int_{\psi-\pi/2}^{\psi-\pi/2+\theta} (2\pi\sin(\phi))^{-1}(\cos^2(t)/\sin^2(\phi) + \sin^2(t))^{-1} dt .
\end{aligned}$$

With $\theta \in [0, \pi/2]$, we note that the integral with respect to t can be evaluated analytically, leading us to

$$\begin{aligned}
&\mathbb{P}\left[\langle \mathbf{G}^1_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^1_{\text{ort}}, \mathbf{y}\rangle < 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{x}\rangle > 0, \langle \mathbf{G}^2_{\text{ort}}, \mathbf{y}\rangle < 0\right] \\
&= \int_{\pi/2-\theta}^{\pi/2} \frac{\mathrm{d}\psi}{2\pi} \int_0^{\pi/2} f(\phi)\mathrm{d}\phi\, \frac{1}{2\pi} \left(\arctan(\tan(\psi - \pi/2 + \theta)\sin(\phi)) - \arctan(\tan(\psi - \pi/2)\sin(\phi))\right) \\
&\leq \int_{\pi/2-\theta}^{\pi/2} \frac{\mathrm{d}\psi}{2\pi} \int_0^{\pi/2} f(\phi)\mathrm{d}\phi\, \frac{\theta}{2\pi} \\
&= \left(\frac{\theta}{2\pi}\right)^2 .
\end{aligned}$$

To deal with $\theta \in [\pi/2, \pi]$, we note that if the angle θ between $\mathbf{x}$ and $\mathbf{y}$ is obtuse, then the angle between $\mathbf{x}$ and $-\mathbf{y}$ is $\pi - \theta$ and therefore acute. Recalling from our definition that $\mathcal{A}_m = \{\operatorname{sign}\left(\langle \mathbf{G}^i_{\text{ort}}, \mathbf{x}\rangle\right) \operatorname{sign}\left(\langle \mathbf{G}^i_{\text{ort}}, \mathbf{y}\rangle\right) = -1\}$, if we denote the corresponding quantity for the pair of vecors $\mathbf{x}$, $-\mathbf{y}$ by $\bar{\mathcal{A}}_m = \{\operatorname{sign}\left(\langle \mathbf{G}^i_{\text{ort}}, \mathbf{x}\rangle\right) \operatorname{sign}\left(\langle \mathbf{G}^i_{\text{ort}}, -\mathbf{y}\rangle\right) = -1\}$, then we in fact have $\bar{\mathcal{A}}_m = \mathcal{A}_m^c$. Therefore, applying the result to the pair of vectors $\mathbf{x}$ and $-\mathbf{y}$ (which have acute angle $\pi - \theta$ between them) and using the inclusion-exclusion principle, we obtain:

$$\begin{aligned}
\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2) &= 1 - \mathbb{P}(\mathcal{A}_1^c) - \mathbb{P}(\mathcal{A}_2^c) + \mathbb{P}(\mathcal{A}_1^c \cap \mathcal{A}_2^c) \\
&< 1 - \mathbb{P}(\mathcal{A}_1^c) - \mathbb{P}(\mathcal{A}_2^c) + \left(\frac{\pi - \theta}{\pi}\right)^2 \\
&= 1 - 2\left(\frac{\pi - \theta}{\pi}\right) + \left(\frac{\pi - \theta}{\pi}\right)^2 \\
&= \left(\frac{\theta}{\pi}\right)^2
\end{aligned}$$

as required.

The second inequality of Proposition 9.3 follows from the inclusion-exclusion principle and the first inequality:

$$\begin{aligned}
\mathbb{P}\left[\mathcal{A}_1^c \cap \mathcal{A}_2^c\right] &= 1 - \mathbb{P}\left[\mathcal{A}_1\right] - \mathbb{P}\left[\mathcal{A}_2\right] + \mathbb{P}\left[\mathcal{A}_1 \cap \mathcal{A}_2\right] \\
&< 1 - \mathbb{P}\left[\mathcal{A}_1\right] - \mathbb{P}\left[\mathcal{A}_2\right] + \left(\frac{\theta}{\pi}\right)^2
\end{aligned}$$

$$= (1 - \mathbb{P}[\mathcal{A}_1])(1 - \mathbb{P}[\mathcal{A}_2])$$
$$= \left(1 - \frac{\theta}{\pi}\right)^2 .$$

□

9.3 Proof of Theorem 4.4

Proof. We will consider the following setting. Given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, each of them is transformed by the nonlinear mapping: $\phi^{\mathbf{M}} : \mathbf{z} \rightarrow \frac{1}{\sqrt{k}}\mathrm{sgn}(\mathbf{Mz})$, where $\mathbf{M} \in \mathbb{R}^{m \times n}$ is some linear transformation and $\mathrm{sgn}(\mathbf{v})$ stands for a vector obtained from $\mathbf{v}$ by applying pointwise nonlinear mapping $\mathrm{sgn} : \mathbb{R} \rightarrow \mathbb{R}$ defined as follows: $\mathrm{sgn}(x) = +1$ if $x > 0$ and $\mathrm{sgn}(x) = -1$ otherwise. The angular distance θ between $\mathbf{x}$ and $\mathbf{y}$ is estimated by: $\hat{\theta}^{\mathbf{M}} = \frac{\pi}{2}(1 - \phi^{\mathbf{M}}(\mathbf{x})^\top \phi^{\mathbf{M}}(\mathbf{y}))$. We will derive the formula for the $\mathrm{MSE}(\hat{\theta}^{\mathbf{M}}(\mathbf{x}, \mathbf{y}))$. One can easily see that the MSE of the considered in the statement of the theorem angular kernel on vectors $\mathbf{x}$ and $\mathbf{y}$ can be obtained from this one by multiplying by $\frac{4}{\pi^2}$.

Denote by $\mathbf{r}^i$ the i^{th} row of $\mathbf{M}$. Notice first that for any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with angular distance θ, the event $E_i = \{\mathrm{sgn}((\mathbf{r}^i)^\top \mathbf{x}) \neq \mathrm{sgn}((\mathbf{r}^i)^\top \mathbf{y})\}$ is equivalent to the event $\{\mathbf{r}^i_{proj} \in \mathcal{R}\}$, where $\mathbf{r}^i_{proj}$ stands for the projection of $\mathbf{r}^i$ into the $\mathbf{x} - \mathbf{y}$ plane and $\mathcal{R}$ is a union of two cones in the $\mathbf{x}$-$\mathbf{y}$ plane obtained by rotating vectors $\mathbf{x}$ and $\mathbf{y}$ by $\frac{\pi}{2}$. Denote $\mathcal{A}^i = \{\mathbf{r}^i_{proj} \in \mathcal{R}\}$ for $i = 1, ..., k$ and $\delta_{i,j} = \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}^j] - \mathbb{P}[\mathcal{A}^i]\mathbb{P}[\mathcal{A}^j]$.

For a warmup, let us start our analysis for the standard unstructured Gaussian estimator case. It is a well known fact that this is an unbiased estimator of θ. Thus

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}^{\mathbf{G}}(\mathbf{x}, \mathbf{y})) = Var(\frac{\pi}{2}(1 - \phi^{\mathbf{M}}(\mathbf{x})^\top \phi^{\mathbf{M}}(\mathbf{y}))) &= \frac{\pi^2}{4} Var(\phi^{\mathbf{M}}(\mathbf{x})^\top \phi^{\mathbf{M}}(\mathbf{y}))) \\ &= \frac{\pi^2}{4}\frac{1}{m^2} Var(\sum_{i=1}^m X_i), \end{aligned} \tag{31}$$

where $X_i = \mathrm{sgn}((\mathbf{r}^i)^\top \mathbf{x})\mathrm{sgn}((\mathbf{r}^i)^\top \mathbf{y})$.

Since the rows of $\mathbf{G}$ are independent, we get

$$Var(\sum_{i=1}^m X_i) = \sum_{i=1}^m Var(X_i) = \sum_{i=1}^m (\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2). \tag{32}$$

From the unbiasedness of the estimator, we have: $\mathbb{E}[X_i] = (-1) \cdot \frac{\theta}{\pi} + 1 \cdot (1 - \frac{\theta}{\pi})$. Thus we get:

$$\mathrm{MSE}(\hat{\theta}^{\mathbf{G}}(\mathbf{x}, \mathbf{y})) = \frac{\pi^2}{4}\frac{1}{m^2}\sum_{i=1}^m (1 - (1 - \frac{2\theta}{\pi})^2) = \frac{\theta(\pi - \theta)}{m}. \tag{33}$$

Multiplying by $\frac{4}{\pi^2}$, we obtain the proof of Lemma 4.2.

Now let us switch to the general case. We first compute the variance of the general estimator $\mathcal{E}$ using matrices $\mathbf{M}$ (note that in this setting we do not assume that the estimator is necessarily unbiased).

By the same analysis as before, we get:

$$Var(\mathcal{E}) = Var(\frac{\pi}{2}(1 - \phi(\mathbf{x})^\top \phi(\mathbf{y}))) = \frac{\pi^2}{4} Var(\phi(\mathbf{x})^\top \phi(\mathbf{y}))) = \frac{\pi^2}{4}\frac{1}{m^2} Var(\sum_{i=1}^m X_i), \tag{34}$$

This time however different X_is are not uncorrelated. We get

$$Var(\sum_{i=1}^{m} X_i) = \sum_{i=1}^{m} Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j) =$$
$$\sum_{i=1}^{m} \mathbb{E}[X_i^2] - \sum_{i=1}^{m} \mathbb{E}[X_i]^2 + \sum_{i \neq j} \mathbb{E}[X_i X_j] - \sum_{i \neq j} \mathbb{E}[X_i]\mathbb{E}[X_j] = \tag{35}$$
$$m + \sum_{i \neq j} \mathbb{E}[X_i X_j] - \sum_{i,j} \mathbb{E}[X_i]\mathbb{E}[X_j]$$

Now, notice that from our previous observations and the definition of $\mathcal{A}^i$, we have

$$\mathbb{E}[X_i] = -\mathbb{P}[\mathcal{A}^i] + \mathbb{P}[\mathcal{A}_c^i], \tag{36}$$

where $\mathcal{A}_c^i$ stands for the complement of $\mathcal{A}^i$.

By the similar analysis, we also get:

$$\mathbb{E}[X_i X_j] = \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}^j] + \mathbb{P}[\mathcal{A}_c^i \cap \mathcal{A}_c^j] - \mathbb{P}[\mathcal{A}_c^i \cap \mathcal{A}^j] - \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}_c^j] \tag{37}$$

Thus we obtain

$$\begin{aligned} Var(\sum_{i=1}^{m} X_i) = m + \sum_{i \neq j} (\mathbb{P}[\mathcal{A}^i \cap \mathcal{A}^j] + \mathbb{P}[\mathcal{A}_c^i \cap \mathcal{A}_c^j] - \mathbb{P}[\mathcal{A}_c^i \cap \mathcal{A}^j] - \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}_c^j] \\ -(\mathbb{P}[\mathcal{A}_c^i] - \mathbb{P}[\mathcal{A}^i])(\mathbb{P}[\mathcal{A}_c^j] - \mathbb{P}[\mathcal{A}^j])) \\ - \sum_i (\mathbb{P}[\mathcal{A}_c^i] - \mathbb{P}[\mathcal{A}^i])^2 = m - \sum_i (1 - 2\mathbb{P}[\mathcal{A}^i])^2 \\ + \sum_{i \neq j} (\mathbb{P}[\mathcal{A}^i \cap \mathcal{A}^j] + \mathbb{P}[\mathcal{A}_c^i \cap \mathcal{A}_c^j] - \mathbb{P}[\mathcal{A}_c^i \cap \mathcal{A}^j] - \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}_c^j] + \\ \mathbb{P}[\mathcal{A}_c^i]\mathbb{P}[\mathcal{A}^j] + \mathbb{P}[\mathcal{A}^i]\mathbb{P}[\mathcal{A}_c^j] - \mathbb{P}[\mathcal{A}_c^i]\mathbb{P}[\mathcal{A}_c^j] - \mathbb{P}[\mathcal{A}^i]\mathbb{P}[\mathcal{A}^j]) \\ = m - \sum_i (1 - 2\mathbb{P}[\mathcal{A}^i])^2 + \sum_{i \neq j} (\delta_1(i,j) + \delta_2(i,j) + \delta_3(i,j) + \delta_4(i,j)), \end{aligned} \tag{38}$$

where

- $\delta_1(i,j) = \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}^j] - \mathbb{P}[\mathcal{A}^i]\mathbb{P}[\mathcal{A}^j]$,
- $\delta_2(i,j) = \mathbb{P}[\mathcal{A}_c^i \cap \mathcal{A}_c^j] - \mathbb{P}[\mathcal{A}_c^i]\mathbb{P}[\mathcal{A}_c^j]$,
- $\delta_3(i,j) = \mathbb{P}[\mathcal{A}_c^i]\mathbb{P}[\mathcal{A}^j] - \mathbb{P}[\mathcal{A}_c^i \cap \mathcal{A}^j]$,
- $\delta_4(i,j) = \mathbb{P}[\mathcal{A}^i]\mathbb{P}[\mathcal{A}_c^j] - \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}_c^j]$.

Now note that

$$\begin{aligned} -\delta_4(i,j) = \mathbb{P}[\mathcal{A}^i] - \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}^j] - \mathbb{P}[\mathcal{A}^i]\mathbb{P}[\mathcal{A}_c^j] \\ = \mathbb{P}[\mathcal{A}^i] - \mathbb{P}[\mathcal{A}^i](1 - \mathbb{P}[\mathcal{A}^j]) - \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}^j] \\ = \mathbb{P}[\mathcal{A}^i]\mathbb{P}[\mathcal{A}^j] - \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}^j] = -\delta_1(i,j) \end{aligned} \tag{39}$$

Thus we have $\delta_4(i,j) = \delta_1(i,j)$. Similarly, $\delta_3(i,j) = \delta_1(i,j)$. Notice also that

$$\begin{aligned} -\delta_2(i,j) = (1 - \mathbb{P}[\mathcal{A}^i])(1 - \mathbb{P}[A^j]) - (\mathbb{P}[\mathcal{A}_c^i] - \mathbb{P}[\mathcal{A}_c^i \cap \mathcal{A}^j]) \\ = 1 - \mathbb{P}[\mathcal{A}^i] - \mathbb{P}[\mathcal{A}^j] + \mathbb{P}[\mathcal{A}^i]\mathbb{P}[\mathcal{A}^j] - 1 + \mathbb{P}[\mathcal{A}^i] + \mathbb{P}[\mathcal{A}_c^i \cap \mathcal{A}^j] \\ = \mathbb{P}[\mathcal{A}^i]\mathbb{P}[\mathcal{A}^j] - \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}^j] = -\delta_1(i,j), \end{aligned} \tag{40}$$

therefore $\delta_2(i,j) = \delta_1(i,j)$.

Thus, if we denote $\delta_{i,j} = \delta_1(i,j) = \mathbb{P}[\mathcal{A}^i \cap \mathcal{A}^j] - \mathbb{P}[\mathcal{A}^i]\mathbb{P}[\mathcal{A}^j]$, then we get

$$Var(\sum_{i=1}^{m} X_i) = m - \sum_i (1 - 2\mathbb{P}[A^i])^2 + 4 \sum_{i \neq j} \delta_{i,j}. \tag{41}$$

Thus we obtain

$$Var(\mathcal{E}) = \frac{\pi^2}{4m^2}[m - \sum_i (1 - 2\mathbb{P}[A^i])^2 + 4\sum_{i \neq j} \delta_{i,j}]. \tag{42}$$

Note that $Var(\mathcal{E}) = \mathbb{E}[(\mathcal{E} - \mathbb{E}[\mathcal{E}])^2]$. We have:

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}^{\mathbf{M}}(\mathbf{x}, \mathbf{y})) = \mathbb{E}[(\mathcal{E} - \theta)^2] &= \mathbb{E}[(\mathcal{E} - \mathbb{E}[\mathcal{E}])^2] + \mathbb{E}[(\mathcal{E} - \theta)^2] - \mathbb{E}[(\mathcal{E} - \mathbb{E}[\mathcal{E}])^2] \\ &= Var(\mathcal{E}) + \mathbb{E}[(\mathcal{E} - \theta)^2 - (\mathcal{E} - \mathbb{E}[\mathcal{E}])^2] \\ &= Var(\mathcal{E}) + (\mathbb{E}[\mathcal{E}] - \theta)^2 \end{aligned} \tag{43}$$

Notice that $\mathcal{E} = \frac{\pi}{2}(1 - \frac{1}{m}\sum_{i=1}^m X_i)$. Thus we get:

$$\mathrm{MSE}(\hat{\theta}^{\mathbf{M}}(\mathbf{x}, \mathbf{y})) = \frac{\pi^2}{4m^2}[m - \sum_i (1 - 2\mathbb{P}[A^i])^2 + 4\sum_{i \neq j} \delta_{i,j}] + \frac{\pi^2}{m^2}\sum_i (\mathbb{P}(\mathcal{A}^i) - \frac{\theta}{\pi})^2. \tag{44}$$

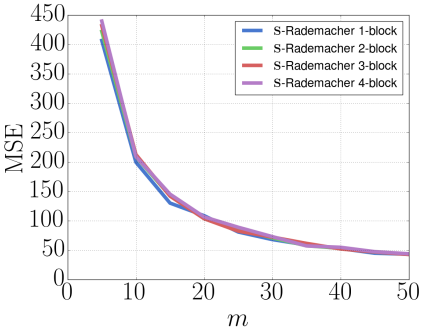
Now it remains to multiply the expression above by $\frac{4}{\pi^2}$ and that completes the proof. □

Remark 9.4. *Notice that if $\mathbb{P}(\mathcal{A}^i) = \frac{\theta}{\pi}$ (this is the case for the standard unstructured estimator as well as for the considered by us estimator using orthogonalized version of Gaussian vectors) and if rows of matrix $\mathbf{M}$ are independent then the general formula for* MSE *for the estimator of an angle reduces to $\frac{(\pi-\theta)\theta}{m}$. If the first property is satisfied but the rows are not necessarily independent (as it is the case for the estimator using orthogonalized version of Gaussian vectors) then whether the* MSE *is larger or smaller than for the standard unstructured case is determined by the sign of the sum $\sum_{i \neq j} \delta_{i,j}$. For the estimator using orthogonalized version of Gaussian vectors we have already showed that for every $i \neq j$ we have: $\delta_{i,j} > 0$ thus we obtain estimator with smaller* MSE*. If $\mathbf{M}$ is a product of blocks $\mathbf{HD}$ then we both have: an estimator with dependent rows and with bias. In that case it is also easy to see that $\mathbb{P}(\mathcal{A}^i)$ does not depend on the choice of i. Thus there exists some ϵ such that $\epsilon = \mathbb{P}(\mathcal{A}^i) - \frac{\theta}{\pi}$. Thus the estimator based on the $\mathbf{HD}$ blocks gives smaller* MSE *iff:*

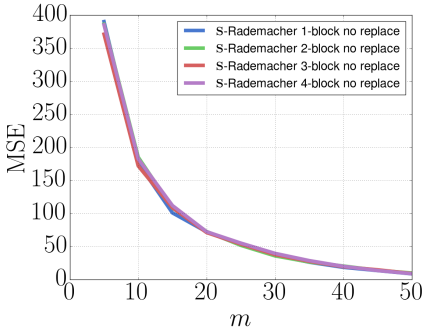
$$\sum_{i \neq j} \delta_{i,j} + m\epsilon^2 < 0.$$

10 Further comparison of variants of OJLT based on SD-product matrices

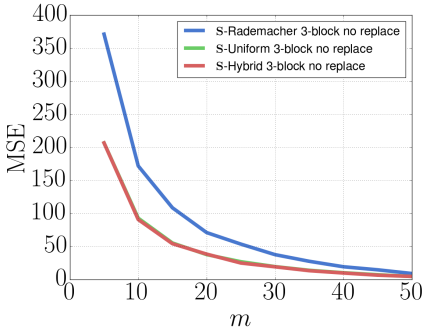
In this section we give details of further experiments complementing the theoretical results of the main paper. In particular, we explore the various parameters associated with the $\mathbf{SD}$-product matrices introduced in §2. In all cases, as in the experiments of §6, we take the structured matrix $\mathbf{S}$ to be the normalized Hadamard matrix $\mathbf{H}$. All experiments presented in this section measure the MSE of the OJLT inner product estimator for two randomly selected data points in the `g50c` data set. The MSE figures are estimated on the basis of $1,000$ repetitions. All results are displayed in Figure 3.



(a) Comparison of estimators based on **S**-Rademacher matrices with a varying number of **SD** matrix blocks, using the with replacement sub-sampling strategy.



(b) Comparison of estimators based on **S**-Rademacher matrices with a varying number of **SD** matrix blocks, using the sub-sampling strategy without replacement.



(c) Comparison of the use of $\mathbf{M}^{(3)}_{\mathbf{S}\mathcal{R}}$, $\mathbf{M}^{(3)}_{\mathbf{S}\mathcal{H}}$, and $\mathbf{M}^{(3)}_{\mathbf{S}\mathcal{U}}$ (introduced in §8.7) for dimensionality reduction. All use sub-sampling without replacement. The curves corresponding to the latter two random matrices are indistinguishable.

Figure 3: Results of experiments comparing OJLTs for a variety of **SD**-matrices.