# Variance Reduction in Sketching Algorithms via Complex Random Variables

**Anonymous authors**
**Paper under double-blind review**

## Abstract

The seminal `CountSketch` algorithm of Charikar et al. (2004) compresses high-dimensional real-valued vectors while approximately preserving pairwise inner products in time proportional to the sparsity of the input data. However, the estimator's high variance limits its reliability. In this work, we propose a simple modification of the `CountSketch` algorithm that not only reduces the variance of the estimate but also maintains its input sparsity running time. Our key idea is to replace real-valued Rademacher $\{-1, 1\}$ used in `CountSketch` with $\{1, \omega, \omega^2, \omega^3\}$(fourth roots of unity). We further extend this idea to the well-known sketching algorithms - `TensorSketch` Pham & Pagh (2013) and `Recursive TensorSketch` Ahle et al. (2020) for a high-degree polynomial kernel, and obtain improvements in the variance. For `TensorSketch`, our proposal achieved exponential improvements in the variance, reducing it from $O(3^p)$ to $O(2^p)$. Wacker et al. (2024) also gives a similar improvement in the variance by exploiting complex random variables. However, the main advantage of our proposal is its running time, which depends on the input sparsity - $O(p \cdot \text{nnz}(\mathbf{x}))$, whereas for Wacker et al. (2024) it is $O(pd)$, for an input $\bigotimes_{i=1}^{p} \mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^d$. We further extend our technique to `Recursive TensorSketch`, a state-of-the-art sketching algorithm for polynomial kernels. Our proposal has lower variance than `Recursive TensorSketch` while retaining the same input sparsity running time.

## 1 Introduction

Dimensionality reduction via randomized linear mappings has become a fundamental algorithmic primitive in large-scale machine learning, enabling computational efficiency while approximately preserving geometric structure such as norms, inner products, and pairwise distances. Classical results such as the JL lemma Johnson & Lindenstrauss (1984) establish that high-dimensional vectors can be embedded into a lower-dimensional space while approximately maintaining pairwise Euclidean distances. Subsequent work has focused on improving the time complexity of such embeddings, most notably through the Fast Johnson–Lindenstrauss Transform (FJLT) Ailon & Chazelle (2006). Hashing-based projection algorithms such as `CountSketch` Charikar et al. (2004) provide similar guarantees on preserving pairwise distances while offering extremely efficient sketches whose time complexity scales only with the number of nonzero entries, making them particularly suitable for high-dimensional sparse datasets. However, for inner-product estimation, `CountSketch` Charikar et al. (2004) exhibits relatively higher variance due to the presence of cross-term contributions.

Kernels provide a powerful mechanism for modelling nonlinear relationships by implicitly mapping inputs into high–dimensional feature spaces. In the case of polynomial kernels, the feature representation is realized as the tensor product $\bigotimes_{i=1}^{p} \mathbf{x_i}$ of vectors $\mathbf{x_1} \in \mathbb{R}^{d_1}, \ldots, \mathbf{x_p} \in \mathbb{R}^{d_p}$, which captures higher-order feature interactions. However, explicitly forming this tensor requires $O(\prod_{i=1}^{p} d_i)$ memory, which becomes infeasible even for moderate values of $p$ or $\{d_i\}_{i=1}^{p}$, making direct computation of tensor feature maps computationally expensive in practice. To address this issue, prior work has developed sketching methods that compute a sketch $\mathbf{S}(\bigotimes_{i=1}^{p} \mathbf{x_i})$ without explicitly constructing the full tensor. These include extensions of both (i) JL-type sketching Kar & Karnick (2012), and (ii) hashing-based `CountSketch` Pham & Pagh (2013; 2025) constructions tailored to tensor products. These methods are referred to as `TensorSketch` in the literature.

Both methods Kar & Karnick (2012); Pham & Pagh (2013) offer similar theoretical guarantees on sketch quality, with variance growing as $O(3^p)$, where $p$ is the tensor order. However, the hashing-based approach of Pham & Pagh (2013) achieves superior computational efficiency, with running time depending only on input sparsity rather than the ambient dimension $d_i$, making it preferable for sparse data.

Recently, Wacker et al. (2024) revisited this variance analysis of JL-type `TensorSketch` Kar & Karnick (2012) and showed that it can be improved to $2^p$ from $3^p$. Their key idea is to replace complex gaussian and complex rademacher random variables, in place of real gaussian and real rademacher random variables. Their analysis demonstrates that complex randomness eliminates a subset of cross terms that contribute to the higher variance. However, as their method is extension of dense JL-type projection Kar & Karnick (2012), whose computational cost scales with the dimension (e.g.,$O(Dd)$ for vectors of dimension $d$ and embedding dimension $D$), makes it computationally expensive for high-dimensional sparse datasets.

In this work, we address the above limitation by proposing complex-valued variants of `CountSketch` and `TensorSketch` that achieve variance improvements comparable to those of the complex JL-type estimator of Wacker et al. (2024), while retaining input sparsity running time of the `CountSketch`-based `TensorSketch` of Pham & Pagh (2013). In particular, our methods provide variance bounds of the same asymptotic order as the complex JL-type construction, but with significantly lower computational cost particularly suited for sparse inputs. In addition, we derive a higher-moment bound on the sketching error. Using this bound, we establish a lower bound on sketching dimension in order to achieve $(\epsilon, \delta)$ approximation guarantee for Complex `CountSketch` and Complex `TensorSketch`. These results further extend to guarantees for approximate matrix multiplication and spectral approximation. Our main contributions are summarized as follows :-

- **Complex `CountSketch` for $\mathbf{x} \in \mathbb{R}^d$.** We introduce a complex-valued version of the classical `CountSketch` algorithm Charikar et al. (2004), in which the random sign function is replaced by a random function whose values are drawn independently and uniformly from the four fourth roots of unity. We prove that the resulting estimator is unbiased and admits a strictly smaller variance than the standard real-valued `CountSketch`. Its time complexity remains proportional to the input sparsity $\text{nnz}(\mathbf{x})$ and also proportional to its real counterpart. Also, the variance of the Complex `CountSketch` scales as $O(1/D)$, whereas its real counterpart scales as $O(2/D)$ due to the cancellation of cross-term contributions. This implies $\text{Var}_{\text{CS-complex}} \leq \text{Var}_{\text{CS-real}}$.

- **Complex `TensorSketch` for $\mathbf{x}^{\otimes p} \in \mathbb{R}^{d^p}$.** Building on the above construction, we propose a complex variant of `TensorSketch` in the style of Pham & Pagh (2013) to sketch input $\mathbf{x}^{\otimes p}$. Our analysis shows that the variance bounds of our proposed sketch scales with $2^p$ as compared to $3^p$ of its real counterpart. This implies $\text{Var}_{\text{TS-complex}} \leq \text{Var}_{\text{TS-real}}$. Also, achieving a running time of $O(p(\text{nnz}(\mathbf{x}) + D \log D))$, which is proportional to Real `TensorSketch`.

- **Complex `Recursive TensorSketch`.** Ahle et al. (2020) proposed a `Recursive TensorSketch` mechanism that makes the variance bound independent of the degree of the polynomial, which is a state-of-the-art algorithm for the task. We extend this line of work by developing a Complex `Recursive TensorSketch` and establishing an improved variance bound, along with a running-time guarantee that remains comparable to the `CountSketch`-based `TensorSketch` of Pham & Pagh (2013). The recursive construction runs in time $O(p(\text{nnz}(\mathbf{x}) + D \log D))$, and its variance satisfies $\text{Var}_{\text{RTS-real}} \leq \frac{3}{D}\|\mathbf{x}\|_2^{2p}\|\mathbf{y}\|_2^{2p}$, while our complex variant achieves $\text{Var}_{\text{RTS-complex}} \leq \frac{2}{D}\|\mathbf{x}\|_2^{2p}\|\mathbf{y}\|_2^{2p}$. This implies $\text{Var}_{\text{RTS-complex}} \leq \text{Var}_{\text{RTS-real}}$, reflecting the improvement obtained by eliminating additional cross-term contributions via complex-valued random function. To the best of our knowledge, there are no existing results that improve the variance bounds of `Recursive TensorSketch` Ahle et al. (2020).

- **Concentration analysis.** We derive a lower bound on the sketching dimension $D = \Omega\left(\max\left\{\varepsilon^{-2}, \varepsilon^{-1}\|\mathbf{x}\|_\infty^2 \log\left(\frac{1}{\delta}\right)\right\} \log\left(\frac{1}{\delta}\right)\right)$ in order to achieve $(\epsilon, \delta)$ guarantee for Complex `CountSketch` and Complex `TensorSketch`. The bound follows from higher moment analysis of the sketching error, adapting the framework of Freksen et al. (2018) to our setting, where the random function $s(\cdot)$ independently and uniformly samples from the fourth roots of unity $\{1, \omega, \omega^2, \omega^3\}$

instead of Rademacher random variables. As corollaries, we obtain bounds for approximate matrix multiplication and spectral approximation.

**Intuition behind our approach:** Our key idea is to employ random function whose values are drawn independently and uniformly from the four fourth roots of unity in the sketching techniques. This is done by replacing real Rademacher signs with a distribution supported on $\{1, \omega, \omega^2, \omega^3\}$, where $\omega$ denotes a fourth root of unity. Choosing the fourth roots of unity results in a reduction in variance due to a crucial property. Specifically, for a random variable uniformly distributed over $\{1, \omega, \omega^2, \omega^3\}$, the expectation of its square is zero, which results in several cross terms vanishing in the moment expansion, leading to a smaller variance.

In the case of real-valued Rademacher $\{-1, 1\}$, where the corresponding second moments are nonzero, which results in a larger variance. In contrast, if $s$ is uniformly distributed over $\{1, \omega, \omega^2, \omega^3\}$(fourth roots of unity), then $\mathbb{E}[s] = 0,\ \mathbb{E}[s^2] = 0,\ \mathbb{E}[|s|^2] = 1$. The property in which $\mathbb{E}[s^2] = 0$ is crucial for the reduction of variance, as it vanishes quadratic cross terms. Although higher roots of unity also satisfy this property, they do not provide further variance reduction instead, they increase randomness by introducing a larger set of random variables. Therefore, the fourth roots of unity are the most suitable choice for achieving variance reduction.

We incorporate this complex randomization into `CountSketch`, `TensorSketch`, and `Recursive TensorSketch` (see Definitions 2,3 and 4), and show that it leads to improved variance behaviour while preserving input–sparsity running time of their real counterparts. The use of complex random variables in sketching has recently attracted renewed attention, with several works demonstrating that complex-valued constructions can yield estimators with reduced variance. In particular, Wacker et al. (2024) introduces a complex JL-type tensor sketch and proves asymptotically smaller variance than its real-valued counterpart. Recent work on implicit matrix trace estimators has also highlighted performance benefits arising from complex random variables Meyer & Avron (2026); Choromanski et al. (2017).

To the best of our knowledge, none of the existing work applies complex random variables to `CountSketch` and its extension `TensorSketch` Pham & Pagh (2013) in order to obtain improved variance guarantees. The analysis of our estimators is technically involved, especially in the recursive setting, since the `Recursive TensorSketch` is constructed by repeatedly combining degree-2 tensor sketches, in a binary-tree structure, which introduces dependent patterns that must be carefully controlled in the variance analysis.

**Implications of our results:** Sketching algorithm such as `CountSketch` Charikar et al. (2004) are widely used for multitask learning Weinberger et al. (2009), low-rank approximation and regression Clarkson & Woodruff (2017) and Compressing neural networks Chen et al. (2015). Our proposed Complex `CountSketch` retains the same asymptotic time complexity while achieving more accurate results, and can be substituted for the `CountSketch` Charikar et al. (2004).

`TensorSketch` Pham & Pagh (2013) and `Recursive TensorSketch` Ahle et al. (2020) further enable efficient linear SVM training Sun et al. (2018); Li et al. (2019), scalable deep learning and the analysis of over-parameterized neural networks Yehudai & Shamir (2019); Zandieh et al. (2021), compact bilinear pooling for fine grained visual recognition Gao et al. (2016) and multimodal fusion models Fukui et al. (2016). In the cases where all inputs are identical, i.e., $\mathbf{x} := \mathbf{x_1} = \cdots = \mathbf{x_p}$, the tensor product corresponds to the feature map of the degree-$p$ polynomial kernel. Polynomial kernels are widely used in areas such as natural language processing Goldberg & Elhadad (2008), recommender systems Rendle (2010), and genomics Aschard (2016). We consider the polynomial kernel $k(\mathbf{x}, \mathbf{y}) = \big(c + \langle \mathbf{x}, \mathbf{y} \rangle\big)^p, \quad c \geq 0,\ p \in \mathbb{N}$. The constant can be absorbed by augmenting each vector with an additional coordinate $\sqrt{c}$ such as $\tilde{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_d, \sqrt{c}), \tilde{\mathbf{y}} = (\mathbf{y}_1, \ldots, \mathbf{y}_d, \sqrt{c})$, providing the homogeneous form $k(\mathbf{x}, \mathbf{y}) = (\tilde{\mathbf{x}}^\top \tilde{\mathbf{y}})^p = (\tilde{\mathbf{x}}^{\otimes p})^\top \tilde{\mathbf{y}}^{\otimes p}$. Thus, approximating the tensor features $\tilde{\mathbf{x}}^{\otimes p}$ and $\tilde{\mathbf{y}}^{\otimes p}$ via `TensorSketch` or `Recursive TensorSketch` and their inner product provides an efficient estimation of the polynomial kernel. Our proposed Complex `TensorSketch` and Complex `Recursive TensorSketch` provide more accurate estimates than their real-valued counterparts while maintaining the same asymptotic time complexity, and therefore serves as better alternatives for above applications.

**Organization of the paper:** The remainder of this paper is organized as follows. Section 2 reviews related work on randomized sketching, polynomial kernel approximation, and variance reduction techniques. Section 3 introduces the necessary preliminaries, including `CountSketch`, `TensorSketch`, `Recursive`

`TensorSketch`, and polynomial kernels. In Section 4, we present our main theoretical results, developing complex-valued variants of `CountSketch`, `TensorSketch`, and `Recursive TensorSketch`, along with proofs of unbiased estimation, variance bounds, time complexity, and concentration analysis. Section 5 reports experimental results and comparisons on both synthetic and real-world datasets. Finally, Section 6 concludes the paper and discusses limitations and directions for future work.

## 2 Related Work

Dimensionality reduction for vectors is grounded in the Johnson–Lindenstrauss (JL) lemma Johnson & Lindenstrauss (1984), which states that any set of $n$ points in $\mathbb{R}^d$ can be embedded into $\mathbb{R}^D$ with distortion $(1 \pm \varepsilon)$ provided $D = O(\varepsilon^{-2} \log n)$, using a random projection computable in $O(dD)$ time. A large body of work has since focused on reducing the computational cost of such embeddings. Examples include very sparse random projections Achlioptas (2003); Li et al. (2006), the Fast JL Transform (FJLT) Ailon & Chazelle (2006), to accelerate projection while preserving JL guarantees Jin et al. (2019).

Along with this, hashing-based methods such as `CountSketch` Charikar et al. (2004) provide embeddings computable in $O(\mathrm{nnz}(\mathbf{x}))$ time, making them particularly effective for high-dimensional sparse datasets, and streaming regimes. For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the real-valued `CountSketch` estimator has variance $\mathrm{Var}_{\text{CS-real}} = \frac{1}{D}\big(\langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - 2\sum_i x_i^2 y_i^2\big)$, where the additional correlation term $\langle \mathbf{x}, \mathbf{y}\rangle^2$ together with $-\sum_i x_i^2 y_i^2$ arises from the use of real Rademacher signs. We propose a simple variant of the `CountSketch` algorithm, in which the (rademacher) sign random variable is replaced by the fourth root of unity of a complex random variable. This small change eliminates these cross-term interactions and leads to an improved variance $\mathrm{Var}_{\text{CS-complex}} = \frac{1}{D}\big(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \sum_i x_i^2 y_i^2\big)$ while preserving exactly the same input-sparsity running time. Notably, this variance coincides with that achieved by complex JL-type estimators based on dense random projections Wacker et al. (2024). However, their proposal Wacker et al. (2024) requires $O(Dd)$ time due to matrix–vector multiplication, whereas our method attains the same variance bounds at lower computational cost, which is similar to its real counterpart, i.e. $O(\mathrm{nnz}(\mathbf{x}) + \mathrm{nnz}(\mathbf{y}))$ time.

The above methods were extended to give sketching algorithms for polynomial kernels, in which feature representation is realized as the tensor product $\bigotimes_{i=1}^p \mathbf{x_i}$ of vectors $\mathbf{x_1} \in \mathbb{R}^{d_1}, \ldots, \mathbf{x_p} \in \mathbb{R}^{d_p}$ to capture higher order feature interactions. Explicitly forming this tensor requires $O(\prod_{i=1}^p d_i)$ memory, which becomes infeasible even for moderate values of $p$ or $\{d_i\}_{i=1}^p$, making direct computation of tensor feature maps computationally expensive in practice. To address this issue, prior work has developed sketching methods that compute a sketch $\mathbf{S}(\bigotimes_{i=1}^p \mathbf{x_i})$ without explicitly constructing the full tensor. These include extensions of JL-type embeddings Kar & Karnick (2012), as well as hashing-based `CountSketch` Pham & Pagh (2013) constructions tailored to tensor products – referred to as `TensorSketch` in the literature. More precisely, for vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, both `TensorSketch` estimators satisfies the variance expression $\mathrm{Var}_{\text{TS-real}} = \frac{1}{D}\big(\big(2\langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - 2\sum_i x_i^2 y_i^2\big)^p - \langle \mathbf{x}, \mathbf{y}\rangle^{2p}\big)$, where the additional multiplicative factor of 2 in front of $\langle \mathbf{x}, \mathbf{y}\rangle^2$ reflects the correlation structure induced by real Rademacher signs and leads to a variance that grows at a $3^p$-type rate. Although the estimators associated with both methods Kar & Karnick (2012); Pham & Pagh (2013) exhibit variance scaling on the order of $\Theta(3^p/D)$ for degree $p$ polynomial features, the hashing-based method of Pham & Pagh (2013) achieves superior computational efficiency. Its running time depends solely on input $d_i$, making it the preferred choice over Kar & Karnick (2012) for sparse data.

Recent work by Wacker et al. (2024) demonstrates that replacing real rademacher/gaussian random variables with complex-valued ones improves this dependence to order $2^p$, i.e., $\mathrm{Var}_{\text{JL-real}} \approx 3^p/D$ versus $\mathrm{Var}_{\text{JL-complex}} \approx 2^p/D$. However, their construction relies on dense JL-style projections having time complexity $O(Dd)$, which makes the approach significantly less efficient for high dimensional sparse datasets despite its improved variance behaviour. We propose a sample variant `TensorSketch` Pham & Pagh (2013) in which we replace real rademacher random variables with fourth–root–of–unity complex random variables. This simple trick cancel the cross-term contributions and yield the improved variance $\mathrm{Var}_{\text{TS-complex}} = \frac{1}{D}\big(\big(\langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \sum_i x_i^2 y_i^2\big)^p - \langle \mathbf{x}, \mathbf{y}\rangle^{2p}\big)$ while retaining the same input–sparsity running time $O(p(\mathrm{nnz}(\mathbf{x}) + \mathrm{nnz}(\mathbf{y}) + D\log D))$. Notably, this variance behaviour matches that of complex JL-type tensor feature estimators, where the complex Rademacher formulation satisfies $\mathrm{Var}_{\text{JL-complex-Rad}} = \frac{1}{D}\big(\big(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + \langle \mathbf{x}, \mathbf{y}\rangle^2 - \sum_k x_k^2 y_k^2\big)^p - \langle \mathbf{x}, \mathbf{y}\rangle^{2p}\big)$ and the complex Gaussian formulation yields $\mathrm{Var}_{\text{JL-complex-Gauss}} = \frac{1}{D}\big(\big(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + \langle \mathbf{x}, \mathbf{y}\rangle^2\big)^p - \langle \mathbf{x}, \mathbf{y}\rangle^{2p}\big)$

Wacker et al. (2024). However, these JL-type constructions require dense matrix–vector multiplication with time complexity $O(pDd)$, whereas our Complex `TensorSketch` achieves a comparable $2^p$-type variance dependence while operating in input–sparsity time.

A further line of related work, due to Ahle et al. (2020), aims to mitigate the exponential dependence on the polynomial degree in the variance of the estimator. Their algorithm, referred to as `Recursive TensorSketch`, reduces the variance scaling from exponential to polynomial in $p$. In particular, the variance satisfies $\mathrm{Var}_{\text{RTS-real}} \leq \frac{3}{D}\|\mathbf{x}\|_2^{2p}\|\mathbf{y}\|_2^{2p}$with a running time of $O(p(\mathrm{nnz}(\mathbf{x}) + D\log D))$. Their construction is based on a hierarchical tree composition of degree-2 sketches which continues to rely on real Rademacher signs, and consequently the variance constants retain cross-term effects. In our work, we introduced Complex `Recursive TensorSketch` which uses fourth–root–of–unity complex random variables instead of real rademacher random variables within the recursive composition. Consequently, our proposal retains the same asymptotic running time while achieving variance behaviour of the form $\mathrm{Var}_{\text{RTS-complex}} \leq \frac{2}{D}\|\mathbf{x}\|_2^{2p}\|\mathbf{y}\|_2^{2p}$, smaller than the variance of Ahle et al. (2020). This demonstrates that complex randomization offers a systematic mechanism for improving variance properties in both non-recursive and recursive tensor sketching frameworks.

Variance reduction in the context of randomized sketching algorithms has also been explored extensively through *statistical variance reduction techniques*, such as control variates and maximum–likelihood–based estimation. Representative examples include the use of control variates for improving, compressed matrix multiplication Verma et al. (2025), variance reduction in feature hashing via MLE and control variates Verma et al. (2022), control–variate based improvements for frequency estimators, and sketching algorithms such as Tug–of–War sketches Pratap & Kulkarni (2021); Pratap et al. (2021); Kang et al. (2021). Related ideas have also been explored in the context of sign random projections by incorporating auxiliary information to reduce estimator variance Kang & Wong (2018). While these approaches can substantially reduce variance, they typically incur additional computational or algorithmic overhead, such as estimating correction coefficients, solving auxiliary optimization problems, or designing problem-specific control variates. In contrast, the methods proposed in this work achieve variance reduction through a minimal and principled modification of existing sketching algorithms, namely by replacing real–valued random signs with carefully chosen complex–valued random variables. This modification preserves the original sketching structure and input–sparsity running time, while yielding systematic and provable improvements in variance. As a result, our approach provides a simple alternative to statistical variance reduction techniques, combining strong theoretical guarantees with practical efficiency and ease of integration into existing sketching frameworks.

## 3 Preliminaries

**Notation:** We use the following notation throughout the paper. Bold lowercase letters (e.g., $\mathbf{x}, \mathbf{y}$) denote vectors, bold uppercase letters (e.g., $\mathbf{X}, \mathbf{Y}$) denote matrices, and bold calligraphic letters (e.g., $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}$) denote higher–order tensors. For a positive integer $D$, we write $[D] := \{1, 2, \ldots, D\}$ for the corresponding index set. The sets $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Z}$ denote the real, complex, and integer domains, respectively, and $\boldsymbol{\Sigma}$ denotes a covariance matrix. For any vector $\mathbf{x}$, $\mathrm{nnz}(\mathbf{x})$ denotes the number of its nonzero entries. The symbol $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the standard inner product, $\|\mathbf{x}\|_2$ the Euclidean norm, and $\|\mathbf{X}\|_F$ the Frobenius norm of a matrix. For a complex number $z \in \mathbb{C}$ written as $z = a + ib$, its complex conjugate is defined as $\overline{z}$. The norm of $z$ is $|z| = \sqrt{a^2 + b^2} = \sqrt{z\,\overline{z}}$. The Kronecker product is written as $\otimes$, while "$\cdot$" denotes standard matrix multiplication and the element-wise multiplication is written as "$\circ$". The indicator function is denoted by $\mathbf{1}[\cdot]$. The probabilistic quantities use $\mathbb{E}[\cdot]$ for expectation, $\mathrm{Var}[\cdot]$ for variance, and $\mathrm{Cov}[\cdot]$ for covariance. Further, in the rest of the paper, we refer Real `CountSketch` as Charikar et al. (2004), Real `TensorSketch` as Pham & Pagh (2013), and Real `Recursive TensorSketch` as Ahle et al. (2020), which are stated in Definitions 2, 3, and 4, respectively.

**Definition 1** (Universal Hash Function Cormen et al. (2009) )**.** Let $[d] = \{1, \ldots, d\}$ and $[D] = \{1, \ldots, D\}$. A family of hash functions $\mathcal{H}$ with mappings $h : [d] \to [D]$ is called *universal* if for all distinct $i, j \in [d]$,

$$\Pr_{h\sim\mathcal{H}}\big[h(i) = h(j)\big] \;\leq\; \frac{1}{D}.$$

Equivalently, the collision probability between any two inputs is no larger than that of a uniformly random hash function.

**Definition 2** (`CountSketch` Charikar et al. (2004))**.** Given an input vector $\mathbf{y} \in \mathbb{R}^d$, the `CountSketch` is a randomized linear map $\mathbf{T} \in \mathbb{R}^{D\times d}$ that maps $\mathbf{y}$ to a lower-dimensional vector $\mathbf{z} = \mathbf{T}\,\mathbf{y} \in \mathbb{R}^D$. The `CountSketch` matrix $\mathbf{T}$ is constructed by two hash functions: (a) $h\colon [d] \to [D]$ a 3-wise independent hash function, and (b) $s : [d] \to \{1, -1\}$ a 4-wise independent random sign function. The $j^{th}$ entry of vector $\mathbf{z} \in \mathbb{R}^D$ is computed as

$$z_j = \sum_{h(i)=j} s(i)\, y_i, \qquad \forall j \in \{1, \ldots, D\}.$$

The time complexity of computing the `CountSketch` is $O(\text{nnz}(\mathbf{y}))$, which in the worst case can be $O(d)$.

For pairwise inner product analysis, let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and let $\hat{k}(\mathbf{x}, \mathbf{y}) := \langle \mathbf{Tx}, \mathbf{Ty} \rangle$ denote the CountSketch inner–product estimator. Then

$$\mathbb{E}\Big[\hat{k}(\mathbf{x}, \mathbf{y})\Big] = \langle \mathbf{x}, \mathbf{y} \rangle,$$

and the variance satisfies

$$\text{Var}\Big[\hat{k}(\mathbf{x}, \mathbf{y})\Big] \;=\; \frac{1}{D}\Big(\langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 - 2\sum_i x_i^2 y_i^2\Big).$$

**Definition 3** (`TensorSketch` of Degree $p$ Pham & Pagh (2013))**.** Let $h_1, \ldots, h_p : [d] \to [D]$ be 3–wise independent hash functions, and $\sigma_1, \ldots, \sigma_p : [d] \to \{-1, +1\}$ be 4–wise independent random sign functions. The `TensorSketch` matrix $\mathbf{S} \in \mathbb{R}^{D\times d^p}$ is defined for $r \in [D]$ and $(i_1, \ldots, i_p) \in [d]^p$ as

$$S_{r,(i_1,\ldots,i_p)} = \left(\prod_{t=1}^{p} \sigma_t(i_t)\right) \mathbf{1}\left[\sum_{t=1}^{p} h_t(i_t) \equiv r \pmod{D}\right].$$

For any $\mathbf{x} \in \mathbb{R}^d$, the sketch $\mathbf{S}(\mathbf{x}^{\otimes p})$ can be computed in time $O(p(\text{nnz}(\mathbf{x}) + D \log D))$ using FFT-based convolution.

For pairwise inner product anaysis, let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and define

$$\hat{k}(\mathbf{x}, \mathbf{y}) := \left\langle \mathbf{S}(\mathbf{x}^{\otimes p}), \mathbf{S}(\mathbf{y}^{\otimes p}) \right\rangle,$$

Then, `TensorSketch` provides an unbiased estimator of the degree–$p$ polynomial kernel

$$\mathbb{E}\Big[\hat{k}(\mathbf{x}, \mathbf{y})\Big] = \langle \mathbf{x}, \mathbf{y} \rangle^p,$$

and its variance satisfies

$$\begin{aligned}\text{Var}\Big[\hat{k}(\mathbf{x}, \mathbf{y})\Big] \;&=\; \frac{1}{D}\left(\left(2\langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - 2\sum_{i=1}^{d} x_i^2 y_i^2\right)^p - \langle \mathbf{x}, \mathbf{y} \rangle^{2p}\right), \\ &\leq \frac{3^p - 1}{D} \, \|\mathbf{x}\|_2^{2p} \, \|\mathbf{y}\|_2^{2p}.\end{aligned}$$

**Definition 4** (`Recursive TensorSketch` Ahle et al. (2020))**.** In this work, we refer to the following construction as `Recursive TensorSketch`. We note that the terminology differs slightly from that used in Ahle et al. (2020), where the same sketching framework is presented under a different naming convention.
Given a vector $\mathbf{z} \in \mathbb{R}^{d^p}$ where $p$ takes values that are powers of two, then the `Recursive TensorSketch` is a randomized linear map

$$\Pi^p : \mathbb{R}^{d^p} \to \mathbb{R}^D, \text{defined as,} \;\; \Pi^p := Q^p \cdot T^p, \;\; \text{where}$$

- $T^p = T_1 \otimes T_2 \otimes \cdots \otimes T_p$, with each $T_i \in \mathbb{R}^{D\times d} \quad \forall i \in [p]$ a `CountSketch` matrix (Definition 2),

- $Q^p = S^2 \cdot S^4 \cdots S^{p/2} \cdot S^p$, with each $S^\ell \in \mathbb{R}^{D^{\ell/2} \times D^\ell}$ a Kronecker product of matrices $S_j^\ell \in \mathbb{R}^{D \times D^2}$,
- Each $S_j^\ell$ is a `TensorSketch` matrix where $p = 2$ (Definition 3), and $S^\ell = S_1^\ell \otimes S_2^\ell \otimes \cdots \otimes S_{\ell/2}^\ell$.

The time complexity required to compute the `Recursive TensorSketch` for $\mathbf{z} \in \mathbb{R}^{d^p}$ is $O(p(\mathrm{nnz}(\mathbf{x}) + D \log D))$.

Let $\Pi^p$ be the Recursive TensorSketch map defined above, and for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ defining the estimator

$$\hat{k}(\mathbf{x}, \mathbf{y}) := \left\langle \Pi^p(\mathbf{x}^{\otimes p}), \Pi^p(\mathbf{y}^{\otimes p}) \right\rangle.$$

Then, the `Recursive TensorSketch` provides unbiased estimation as

$$\mathbb{E}\Big[\hat{k}(\mathbf{x}, \mathbf{y})\Big] = \langle \mathbf{x}, \mathbf{y} \rangle^p.$$

Moreover, its variance bound is given as

$$\mathrm{Var}\Big[\hat{k}(\mathbf{x}, \mathbf{y})\Big] \;\le\; \frac{3}{D} \|\mathbf{x}\|_2^{2p} \|\mathbf{y}\|_2^{2p}.$$

**Definition 5** (Polynomial Kernel Schölkopf & Smola (2001))**.** We consider polynomial kernels of degree $p \in \mathbb{N}$ of the form

$$k(\mathbf{x}, \mathbf{y}) = \big(\gamma\, \mathbf{x}^\top \mathbf{y} + \nu\big)^p,$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ with $\gamma, \nu \ge 0$. The parameters $\gamma$ and $\nu$ can be absorbed into the inputs by introducing the augmented vectors

$$\tilde{\mathbf{x}} = \big(\sqrt{\gamma}\, \mathbf{x}^\top,\ \sqrt{\nu}\big)^\top, \qquad \tilde{\mathbf{y}} = \big(\sqrt{\gamma}\, \mathbf{y}^\top,\ \sqrt{\nu}\big)^\top \in \mathbb{R}^{d+1},$$

With this transformation, the kernel admits a homogeneous representation,

$$\big(\gamma\, \mathbf{x}^\top \mathbf{y} + \nu\big)^p = (\tilde{\mathbf{x}}^\top \tilde{\mathbf{y}})^p = \big(\tilde{\mathbf{x}}^{\otimes p}\big)^\top \tilde{\mathbf{y}}^{\otimes p}.$$

where $\tilde{\mathbf{x}}^{\otimes p}$ denotes the $p$-fold tensor product of $\tilde{\mathbf{x}}$. Hence, without loss of generality, we may treat the polynomial kernel as a homogeneous kernel in the augmented space $\mathbb{R}^{d+1}$.

**Definition 6** (Complex Polynomial Sketch (Rademacher/Gaussian) Wacker et al. (2024))**.** Let $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y})^p$ be the polynomial kernel of degree $p$ and let $D \in \mathbb{N}$. For $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, D\}$, generate real random vectors $v_{i,j}, w_{i,j} \in \mathbb{R}^d$ satisfying $\mathbb{E}[v_{i,j}] = \mathbb{E}[w_{i,j}] = 0$ and $\mathbb{E}[v_{i,j} v_{i,j}^\top] = \mathbb{E}[w_{i,j} w_{i,j}^\top] = \mathbf{I}_d$. Define complex vectors

$$z_{i,j} = \frac{1}{\sqrt{2}}\big(v_{i,j} + i\, w_{i,j}\big) \in \mathbb{C}^d, \qquad \mathbb{E}[z_{i,j} z_{i,j}^\top] = \mathbf{I}_d,$$

The complex-valued random feature map is

$$\Phi_C(\mathbf{x}) = \frac{1}{\sqrt{D}}\Big[\prod_{i=1}^p z_{i,1}^\top \mathbf{x},\ \ldots,\ \prod_{i=1}^p z_{i,D}^\top \mathbf{x}\Big]^\top \in \mathbb{C}^D,$$

and the corresponding approximate kernel is

$$\hat{k}_C(\mathbf{x}, \mathbf{y}) = \frac{1}{D} \sum_{j=1}^D \prod_{i=1}^p \big(z_{i,j}^\top \mathbf{x}\big)\big(z_{i,j}^\top \mathbf{y}\big), \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

- When the real components $v_{i,j}$ and $w_{i,j}$ used to form the complex vectors $z_{i,j} = \frac{1}{\sqrt{2}}(v_{i,j} + i\, w_{i,j})$ are sampled i.i.d. from the Rademacher distribution $\{\pm 1\}$, the resulting construction is called the *Complex Rademacher sketch*.
- When the real components $v_{i,j}$ and $w_{i,j}$ used to form the complex vectors $z_{i,j} = \frac{1}{\sqrt{2}}(v_{i,j} + i\, w_{i,j})$ are sampled i.i.d. from the standard Gaussian distribution $\mathcal{N}(0, 1)$, the resulting construction is called the *Complex Gaussian sketch*.

**Unbiasedness.** For both Complex Gaussian and Complex Rademacher sketches,

$$\mathbb{E}\Big[\hat{k}_C(\mathbf{x},\mathbf{y})\Big] = (\mathbf{x}^\top \mathbf{y})^p.$$

**Variance.**

- For *Complex Rademacher sketch*

$$\begin{aligned}\mathrm{Var}\Big[\hat{k}_C(\mathbf{x},\mathbf{y})\Big] &= \frac{1}{D}\left(\left(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + (\mathbf{x}^\top\mathbf{y})^2 - \sum_{k=1}^{d} x_k^2 y_k^2\right)^p - (\mathbf{x}^\top\mathbf{y})^{2p}\right), \\ &\leq \frac{2^p}{D}\Big(\|\mathbf{x}\|_2^{2p}\|\mathbf{y}\|_2^{2p}\Big).\end{aligned}$$

- For *Complex Gaussian sketch*

$$\begin{aligned}\mathrm{Var}\Big[\hat{k}_C(\mathbf{x},\mathbf{y})\Big] &= \frac{1}{D}\Big(\Big(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + (\mathbf{x}^\top\mathbf{y})^2\Big)^p - (\mathbf{x}^\top\mathbf{y})^{2p}\Big), \\ &\leq \frac{2^p}{D}\Big(\|\mathbf{x}\|_2^{2p}\|\mathbf{y}\|_2^{2p}\Big).\end{aligned}$$

## 4 Analysis of Complex Sketching Methods

### 4.1 Complex `CountSketch`

In this subsection, we define the Complex `CountSketch` (Definition 4.1) and derive the expected value and variance of its estimator (Theorem 1). We then compare it to the Real `CountSketch` Charikar et al. (2004), formally showing that the complex estimator always attains variance no greater than that of its real counterpart, while preserving the input-sparsity running time.

**Definition 4.1: Complex `CountSketch` Mapping**

Let $\mathbf{x} \in \mathbb{R}^d$. Define a Complex `CountSketch` matrix $\mathbf{C} \in \mathbb{C}^{D\times d}$ with the following two independent functions

- $h : [d] \to [D]$ is an universal hash function that assigns each coordinate independently and uniformly to one of the $D$ buckets, and
- $s : [d] \to \{1, \omega, \omega^2, \omega^3\}$ is a random function whose values are drawn independently and uniformly from the four fourth roots of unity.

The sketch of a vector $\mathbf{x} \in \mathbb{R}^d$ is defined as $\mathbf{Cx} \in \mathbb{R}^D$, whose $j$-th coordinate is given by

$$(\mathbf{Cx})_j \;=\; \sum_{i=1}^{d} s(i)\, x_i\, \mathbf{1}_{h(i)=j}, \qquad \forall j \in [D], \tag{1}$$

Define the complex sketch mapping

$$\Phi_C(\mathbf{x}) := \big((\mathbf{Cx})_1, (\mathbf{Cx})_2, \ldots, (\mathbf{Cx})_D\big) \in \mathbb{C}^D, \tag{2}$$

Then, the estimate of inner product between two input pairs $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ using Complex `Count Sketch` is defined as

$$\hat{k}_C(\mathbf{x},\mathbf{y}) := \Phi_C(\mathbf{x})^T \overline{\Phi_C(\mathbf{y})} \in \mathbb{C}. \tag{3}$$

**Theorem 1** (Unbiasedness, variance, and sketching time of Complex `CountSketch` inner product estimator)**.** *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and let $\mathbf{C} \in \mathbb{C}^{D\times d}$ denote a Complex `CountSketch` matrix as defined in Definition 4.1. Let*

*$\hat{k}_C(\mathbf{x}, \mathbf{y})$ denote the inner product estimator defined in Definition 4.1. Then the estimator is unbiased and satisfies*

$$\mathbb{E}\Big[\hat{k}_C(\mathbf{x}, \mathbf{y})\Big] = \langle \mathbf{x}, \mathbf{y} \rangle, \tag{4}$$

$$\mathrm{Var}\Big[\hat{k}_C(\mathbf{x}, \mathbf{y})\Big] = \frac{1}{D}\left( \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2 \right). \tag{5}$$

*Moreover, the sketch $\mathbf{Cx}$ and $\mathbf{Cy}$ can be computed in $O(\mathrm{nnz}(\mathbf{x}))$ and $O(\mathrm{nnz}(\mathbf{y}))$ time, respectively.*

*Proof.* We first outline the structure of the proof. To prove unbiasedness, we begin by expanding the inner product expression of the sketched vectors obtained from Complex `CountSketch`. We compute the expectation using the independence of the hash function and complex random hash funtions along with the moment properties $\mathbb{E}[s(i)\overline{s(r)}] = 0$ for $i \neq r$ and $\mathbb{E}[|s(i)|^2] = 1$ due to which all cross terms vanish and we obtain the unbiased estimation of actual inner product. For the variance, we expand the second moment of the estimator and analyze the non-zero terms. Since $\mathbb{E}[s(i)^2] = \mathbb{E}[\overline{s(i)}^2] = 0$ and $\mathbb{E}[s(i)\overline{s(i)}] = \mathbb{E}[|s(i)|^2] = 1$ for all $i \in [d]$ imply that all terms vanish except those corresponding to index configurations with pairwise matchings. Combining these contributions provides a closed-form expression for the second moment, and subtracting the squared mean gives a variance of order $1/D$, completing the proof.

We now provide the detailed argument. By expanding the estimator, we obtain

$$\begin{aligned}
\hat{k}_C(\mathbf{x}, \mathbf{y}) = \Phi_C(\mathbf{x})^T \overline{\Phi_C(\mathbf{y})} &= \langle \mathbf{Cx}, \overline{\mathbf{Cy}} \rangle, \qquad (6) \\
&= \sum_{j=1}^{D} (Cx)_j \, \overline{(Cy)_j}, \\
&= \sum_{j=1}^{D} \Big( \sum_{i=1}^{d} s(i) \mathbf{1}_{h(i)=j} x_i \Big) \Big( \sum_{r=1}^{d} \overline{s(r)} \mathbf{1}_{h(r)=j} y_r \Big), \\
&= \sum_{j=1}^{D} \sum_{i=1}^{d} \sum_{r=1}^{d} s(i)\overline{s(r)} \, \mathbf{1}_{h(i)=j} \, \mathbf{1}_{h(r)=j} \, x_i \, y_r. \qquad (7)
\end{aligned}$$

**Computing Expectation:**
We compute the expected value of Equation (7).

$$\begin{aligned}
\mathbb{E}\Big[\hat{k}_C(\mathbf{x}, \mathbf{y})\Big] &= \sum_{j=1}^{D} \sum_{i,r=1}^{d} x_i y_r \, \mathbb{E}\Big[s(i)\overline{s(r)}\Big] \; \mathbb{E}\big[\mathbf{1}_{h(i)=j} \mathbf{1}_{h(r)=j}\big] \,, \\
&= \sum_{j=1}^{D} \sum_{i=1}^{d} x_i y_i \, \mathbb{E}\big[|s(i)|^2\big] \; \mathbb{E}\Big[\mathbf{1}_{h(i)=j}^2\Big] \; + \; \cdots \\
&\cdots + \sum_{j=1}^{D} \sum_{\substack{i,r=1 \\ i \neq r}}^{d} x_i y_r \, \mathbb{E}\Big[s(i)\overline{s(r)}\Big] \; \mathbb{E}\big[\mathbf{1}_{h(i)=j} \mathbf{1}_{h(r)=j}\big] \,. \qquad (8)
\end{aligned}$$

By independence and symmetry of the functions $h(.)$ and $s(.)$, we have $\mathbb{E}[|s(i)|^2] = 1$ and $\mathbb{E}[s(i)\overline{s(r)}] = 0$ for $i \neq r$. Moreover, since $\mathbf{1}_{h(i)=j}^2 = \mathbf{1}_{\{h(i)=j\}}$ with $h(i)$ uniform on $[D]$,

$$\mathbb{E}[\mathbf{1}_{h(i)=j}^2] = \mathbb{E}[\mathbf{1}_{h(i)=j}] = \frac{1}{D}.$$

Substituting these identities into (8) vanishes cross term and we get

$$\mathbb{E}\Big[\hat{k}_C(\mathbf{x}, \mathbf{y})\Big] = \sum_{j=1}^{D} \frac{1}{D} \sum_{i=1}^{d} x_i y_i = \langle \mathbf{x}, \mathbf{y} \rangle. \tag{9}$$

This completes the proof of unbiasedness. We next turn to the analysis of the variance of the estimator.

**Computing Variance:**
The variance of the complex estimator can be expressed as

$$\mathrm{Var}\Big[\hat{k}_C(\mathbf{x},\mathbf{y})\Big] \;=\; \mathbb{E}\Big[\,\Big|\hat{k}_C(\mathbf{x},\mathbf{y})\Big|^2\,\Big] \;-\; \Big|\mathbb{E}\Big[\hat{k}_C(\mathbf{x},\mathbf{y})\Big]\Big|^2. \tag{10}$$

To evaluate the first term, we expand it using Equation (7) as follows

$$\begin{aligned}\Big|\hat{k}_C(\mathbf{x},\mathbf{y})\Big|^2 = \big|\langle \mathbf{Cx}, \overline{\mathbf{Cy}}\rangle\big|^2 &= \left|\sum_{j=1}^{D}\sum_{i,r=1}^{d} s(i)\,\overline{s(r)}\,\mathbf{1}_{h(i)=j}\mathbf{1}_{h(r)=j}\,x_i y_r\right|^2 \\ &= \sum_{j,j'=1}^{D}\sum_{i,r,p,q=1}^{d} s(i)\,\overline{s(r)}\,\overline{s(p)}\,s(q)\;\mathbf{1}_{h(i)=j}\mathbf{1}_{h(r)=j}\mathbf{1}_{h(p)=j'}\mathbf{1}_{h(q)=j'}x_i y_r x_p y_q. \end{aligned}\tag{11}$$

Taking expectations with respect to the randomness in $h(.)$ and $s(.)$, we obtain

$$\begin{aligned}&\mathbb{E}\left[\Big|\hat{k}_C(\mathbf{x},\mathbf{y})\Big|^2\right] \\ &\quad= \sum_{j,j'=1}^{D}\sum_{i,r,p,q=1}^{d} x_i y_r x_p y_q \mathbb{E}\Big[s(i)\overline{s(r)s(p)}s(q)\Big]\,\mathbb{E}\big[\mathbf{1}_{h(i)=j}\mathbf{1}_{h(r)=j}\mathbf{1}_{h(p)=j'}\mathbf{1}_{h(q)=j'}\big]\,. \end{aligned}\tag{12}$$

We begin with the **case** $j = j'$:

Since the random variables $\{s(i)\}_{i=1}^{d}$ are i.i.d. with $\mathbb{E}[s(i)] = 0$, $\mathbb{E}[|s(i)|^2] = 1$, and $\mathbb{E}[s(i)^2] = 0$, the fourth-order moment

$$\mathbb{E}\Big[s(i)\,\overline{s(r)}\,\overline{s(p)}\,s(q)\Big]$$

is nonzero only when each index appears an even number of times. Following terms which are non-zero:

(a) $i = r = p = q$: $\quad \mathbb{E}[|s(i)|^4] = 1, \qquad \mathbb{E}[\mathbf{1}^4_{h(i)=j}] = \mathbb{E}[\mathbf{1}_{h(i)=j}] = \frac{1}{D}$.

(b) $i = r \neq p = q$: $\quad \mathbb{E}[|s(i)|^2|s(p)|^2] = 1, \qquad \mathbb{E}[\mathbf{1}^2_{h(i)=j}\mathbf{1}^2_{h(p)=j}] = \frac{1}{D^2}$.

(c) $i = p \neq r = q$: $\quad \mathbb{E}[|s(i)|^2|s(r)|^2] = 1, \qquad \mathbb{E}[\mathbf{1}^2_{h(i)=j}\mathbf{1}^2_{h(r)=j}] = \frac{1}{D^2}$.

All other cases are zero. Adding the contributions from the above cases, we obtain

$$\sum_{i=1}^{d} x_i^2 y_i^2 \;+\; \frac{1}{D}\left(\sum_{\substack{i,p=1\\ i\neq p}}^{d} x_i y_i x_p y_p \;+\; \sum_{\substack{i,r=1\\ i\neq r}}^{d} x_i^2 y_r^2\right). \tag{13}$$

We next consider the **case** $j \neq j'$: Since the same index cannot hash to two different buckets, all terms vanish except the following case.

(a) $i = r \neq p = q$: $\;\mathbb{E}[|s(i)|^2|s(p)|^2] = 1, \qquad \mathbb{E}[\mathbf{1}^2_{h(i)=j}\mathbf{1}^2_{h(p)=j'}] = \frac{1}{D^2}$.

Therefore we have,

$$\frac{D-1}{D}\sum_{i\neq p} x_i y_i x_p y_p. \tag{14}$$

Combining Equation (13) and Equation (14), we get

$$\mathbb{E}\left[\left|\hat{k}_C(\mathbf{x},\mathbf{y})\right|^2\right] = \sum_i^d x_i^2 y_i^2 + \frac{1}{D}\left(\sum_{i\neq p} x_i y_i x_p y_p + \sum_{i\neq r} x_i^2 y_r^2\right) + \frac{D-1}{D}\left(\sum_{i\neq p} x_i y_i x_p y_p\right), \tag{15}$$

$$= \langle \mathbf{x},\mathbf{y}\rangle^2 + \frac{1}{D}\sum_{i\neq r} x_i^2 y_r^2. \tag{16}$$

Substituting Equation (16) and Equation (9) into Equation (10), we get

$$\mathrm{Var}\left[\hat{k}_C(\mathbf{x},\mathbf{y})\right] = \langle \mathbf{x},\mathbf{y}\rangle^2 + \frac{1}{D}\left(\sum_{i\neq r} x_i^2 y_r^2\right) - \langle \mathbf{x},\mathbf{y}\rangle^2, \tag{17}$$

$$= \frac{1}{D}\Big(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \sum_i x_i^2 y_i^2\Big). \tag{18}$$

□

**Remark 1** (Sketching time for Complex `CountSketch`)**.** *For a vector $\mathbf{x} \in \mathbb{R}^d$, the Complex* ***`CountSketch`*** *sketch $\mathbf{Cx}$ can be computed in $O(\mathrm{nnz}(\mathbf{x}))$ time. This is because each nonzero entry $x_i$ contributes to exactly one bucket $h(i)$ with a single multiplication by the corresponding complex random variable $s(i)$ and a single addition, while zero entries require no computation.*

**Remark 2** (Comparison between Real Charikar et al. (2004) & Complex `CountSketch`)**.** *The variance of the Complex* ***`CountSketch`*** *estimator is always less than or equal to that of the real estimator. Using the variance formulas*

$$\mathrm{Var}_{\textit{CS-real}} = \frac{1}{D}\left(\|\mathbf{x}\|^2\|\mathbf{y}\|^2 + \langle \mathbf{x},\mathbf{y}\rangle^2 - 2\sum_i \mathbf{x}_i^2\mathbf{y}_i^2\right),$$

$$\mathrm{Var}_{\textit{CS-complex}} = \frac{1}{D}\left(\|\mathbf{x}\|^2\|\mathbf{y}\|^2 - \sum_i \mathbf{x}_i^2\mathbf{y}_i^2\right),$$

*their difference equals*

$$\mathrm{Var}_{\textit{CS-real}} - \mathrm{Var}_{\textit{CS-complex}} = \frac{1}{D}\left(\langle \mathbf{x},\mathbf{y}\rangle^2 - \sum_i \mathbf{x}_i^2\mathbf{y}_i^2\right).$$

*By the Cauchy–Schwarz inequality, $\langle \mathbf{x},\mathbf{y}\rangle^2 \geq \sum_i \mathbf{x}_i^2\mathbf{y}_i^2$, hence*

$$\mathrm{Var}_{\textit{CS-complex}} \leq \mathrm{Var}_{\textit{CS-real}} \quad \textit{always},$$

*with equality only when $\mathbf{x}$ and $\mathbf{y}$ are parallel and non-negative coordinatewise.*

The Table 4.1 summarizes the variance bounds and sketching time of Complex `CountSketch` and other baseline methods.

### 4.2 Complex `TensorSketch`

In this subsection, we define the Complex `TensorSketch` (Definition 4.2) and its corresponding kernel estimate. We then prove unbiasedness of the Complex `TensorSketch`, derive its variance bound, and analyze its sketching time (Theorem 3). We further provide a direct comparison with the Real `TensorSketch` Pham & Pagh (2013), formally showing that the complex tensor estimator always attains variance no larger than that of its real counterpart while retaining its advantage of input sparsity running time.

| Algorithm | Variance | Sketching Time |
|---|---|---|
| `CountSketch`(Complex) | $\frac{1}{D}\left(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \sum_i x_i^2 y_i^2\right)$ | $O(\mathrm{nnz}(\mathbf{x}) + \mathrm{nnz}(\mathbf{y}))$ |
| `CountSketch`(Real)Charikar et al. (2004) | $\frac{1}{D}\left(\langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - 2\sum_i x_i^2 y_i^2\right)$ | $O(\mathrm{nnz}(\mathbf{x}) + \mathrm{nnz}(\mathbf{y}))$ |
| JL(Complex Rademacher)Wacker et al. (2024) | $\frac{1}{D}\left(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \sum_i x_i^2 y_i^2\right)$ | $O(Dd)$ |
| JL(Complex Guassian)Wacker et al. (2024) | $\frac{1}{D}\left(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2\right)$ | $O(Dd)$ |

Table 1: **Comparison of variance bounds and sketching time for Complex `CountSketch` and baseline methods.** Here, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ denote input vectors, $D$ is the sketch dimension. The function $\mathrm{nnz}(\mathbf{x})$ denotes the number of nonzero entries of $\mathbf{x}$. All variance expressions are stated for inner product estimates $\langle \mathbf{x}, \mathbf{y}\rangle$ obtained from these baseline methods.

**Definition 4.2: Complex `TensorSketch` Mapping**

Let $\mathbf{x} \in \mathbb{R}^d$ and fix an integer $p \geq 1$. Define a Complex `TensorSketch` mapping $\mathbf{C} : \mathbb{R}^{d^p} \to \mathbb{C}^D$ constructed from $p$ independent Complex `CountSketch` mappings. For each $r \in [p]$, let

- $h_r : [d] \to [D]$ be an universal hash function that assigns each coordinate independently and uniformly to one of the $D$ buckets, and
- $s_r : [d] \to \{1, \omega, \omega^2, \omega^3\}$ be a random function whose values are drawn independently and uniformly from the four fourth roots of unity.

For each $r \in [p]$, let $\mathbf{C}_r \in \mathbb{C}^{D\times d}$ denote the Complex `CountSketch` matrix induced by $(h_r, s_r)$ as in Definition 4.1. The `TensorSketch` of $\mathbf{x}^{\otimes p} \in \mathbb{R}^{d^p}$ is defined implicitly via convolution of the $p$ sketches, and can be computed efficiently using the Fast Fourier Transform as

$$\mathbf{C}\mathbf{x}^{\otimes p} \;:=\; \mathrm{FFT}^{-1}(\mathrm{FFT}(\mathbf{C}_1\mathbf{x}) \circ \mathrm{FFT}(\mathbf{C}_2\mathbf{x}) \circ \cdots \circ \mathrm{FFT}(\mathbf{C}_p\mathbf{x})) \;\in\; \mathbb{C}^D, \tag{19}$$

where $\circ$ denotes component-wise multiplication.
The corresponding degree-$p$ polynomial kernel estimator is defined as

$$\hat{k}_C(\mathbf{x}, \mathbf{y}) \;:=\; \left\langle \mathbf{C}\mathbf{x}^{\otimes p}, \overline{\mathbf{C}\mathbf{y}^{\otimes p}} \right\rangle \;=\; \left(\mathbf{C}\mathbf{x}^{\otimes p}\right)^T \overline{\mathbf{C}\mathbf{y}^{\otimes p}} \;\in\; \mathbb{C}. \tag{20}$$

The following Theorem 3 establishes the unbiasedness, variance bound, and sketching time of the Complex `TensorSketch` defined in Definition 4.2.

**Theorem 3** (Unbiasedness, variance, and sketching time of Complex `TensorSketch` for degree-$p$ polynomial kernel)**.** *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p} \in \mathbb{R}^{d^p}$. Let $\mathbf{C}$ denotes Complex `TensorSketch` matrix as defined in Definiton 4.2. Then, the degree-p polynomial kernel estimator $\hat{k}_C(\mathbf{x}, \mathbf{y}) := \left\langle \mathbf{C}\mathbf{x}^{\otimes p}, \overline{\mathbf{C}\mathbf{y}^{\otimes p}} \right\rangle$ satisfies*

$$\mathbb{E}\left[\hat{k}_C(\mathbf{x}, \mathbf{y})\right] = \langle \mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle^p, \tag{21}$$

$$\mathrm{Var}\left[\hat{k}_C(\mathbf{x}, \mathbf{y})\right] \leq \frac{1}{D}\left(\left(\langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2 - \sum_{i=1}^d x_i^2 y_i^2\right)^p - \langle \mathbf{x}, \mathbf{y}\rangle^{2p}\right), \tag{22}$$

$$\leq \frac{2^p - 1}{D}\,\|\mathbf{x}\|_2^{2p}\,\|\mathbf{y}\|_2^{2p}. \tag{23}$$

*Moreover, the sketches $\mathbf{C}\mathbf{x}^{\otimes p}$ and $\mathbf{C}\mathbf{y}^{\otimes p}$ can be computed in $O(p\,(\mathrm{nnz}(\mathbf{x}) + D\log D))$ and $O(p\,(\mathrm{nnz}(\mathbf{y}) + D\log D))$ time, respectively.*

*Proof.* We first outline the structure of the proof. Complex `TensorSketch` is viewed as a Complex `CountSketch` applied to the $p$-fold tensor products $\mathbf{x}^{\otimes p}$ and $\mathbf{y}^{\otimes p}$ via suitably defined composite hash and

complex random functions. We prove Unbiasedness by expanding the sketched inner product and using properties of the expected value of the random function $s(.)$ to eliminate all cross terms. To analyze the variance, we expand the second moment of the estimator and using the independence between the functions $(H, S)$, the second moment reduces to a scaled second-moment expression involving only the random function $s(.)$. This expression is bounded using a complex AMS moment bound, proved later in Lemma 6. Finally, the variance bound is simplified using the Cauchy-Schwarz inequality, resulting in an $O(1/D)$ bound.

We now present the detailed proof. We begin by noting that the `TensorSketches` $\mathrm{C}\mathbf{x}^{\otimes p}, \mathrm{C}\mathbf{y}^{\otimes p}$ are the `CountSketches` of the tensor product $X := \mathbf{x}^{\otimes p},\ Y := \mathbf{y}^{\otimes p}$ using the two aggregated functions $H : [d]^p \mapsto [D]$ and $S : [d]^p \to \{1, \omega, \omega^2, \omega^3\}$ such that:

$$H(i_1, \ldots, i_p) = \left( \sum_{j=1}^{p} h_j(i_j) \right) \bmod D, \tag{24}$$

$$S(i_1, \ldots, i_p) = \prod_{j=1}^{p} s_j(i_j). \tag{25}$$

Also note that $H(.)$ is 2-wise independent Pătraşcu & Thorup (2012).

For further proof, we use $u, v \in [d]^p$ as the indices of vectors $X, Y$ of dimension $d^p$. Then we expand $\hat{k}_C(\mathbf{x}, \mathbf{y})$ as,

$$\hat{k}_C(\mathbf{x}, \mathbf{y}) = \langle \mathbf{C}X, \overline{\mathbf{C}Y} \rangle = \sum_{u,v \in [d]^p} X_u\, Y_v\, S(u)\, \overline{S(v)}\, \mathbf{1}_{[H(u)=H(v)]}, \tag{26}$$

$$= \langle X, Y \rangle \ + \sum_{u \neq v} X_u\, Y_v\, S(u)\, \overline{S(v)}\, \mathbf{1}_{[H(u)=H(v)]}. \tag{27}$$

As we know, $\mathbb{E}[S(u)\,\overline{S(v)}] = 0, \forall\ u \neq v$. Then we have

$$\mathbb{E}\Big[\hat{k}_C(\mathbf{x}, \mathbf{y})\Big] = \langle X, Y \rangle = \langle \mathbf{x}, \mathbf{y} \rangle^p. \tag{28}$$

For the variance, we first compute $\mathbb{E}\Big[|\hat{k}_C(\mathbf{x}, \mathbf{y})|^2.\Big]$ Let's first expand the second moment term,

$$|\hat{k}_C(\mathbf{x}, \mathbf{y})|^2 = \langle \mathbf{C}\mathbf{x}^{\otimes p}, \overline{\mathbf{C}\mathbf{y}^{\otimes p}} \rangle \langle \overline{\mathbf{C}\mathbf{x}^{\otimes p}}, \mathbf{C}\mathbf{y}^{\otimes p} \rangle \tag{29}$$

$$= \left( \langle X,Y \rangle + \sum_{u \neq v} X_u Y_v S(u) \overline{S(v)} \mathbf{1}_{[H(u)=H(v)]} \right) \left( \langle X,Y \rangle + \sum_{u \neq v} X_u Y_v \overline{S(u)} S(v) \mathbf{1}_{[H(u)=H(v)]} \right), \tag{30}$$

$$= \langle X, Y \rangle^2 + \langle X, Y \rangle \left( \sum_{u \neq v} X_u Y_v S(u) \overline{S(v)} \mathbf{1}_{[H(u)=H(v)]} + \sum_{u \neq v} X_u Y_v \overline{S(u)} S(v) \mathbf{1}_{[H(u)=H(v)]} \right) + \cdots$$

$$\cdots + \left| \left( \sum_{u \neq v} X_u Y_v S(u) \overline{S(v)} \mathbf{1}_{[H(u)=H(v)]} \right) \right|^2. \tag{31}$$

Now, take the expectation of $|\hat{k}_C(\mathbf{x}, \mathbf{y})|^2$ and we know that $\mathbb{E}\left[\overline{S(u)} S(v)\right] = \mathbb{E}\left[S(u)\overline{S(v)}\right] = 0, \forall u \neq v$. Then,

$$\mathbb{E}\left[ |\langle \mathbf{C}\mathbf{x}^{\otimes p}, \overline{\mathbf{C}\mathbf{y}^{\otimes p}} \rangle|^2 \right] = \langle X, Y \rangle^2 + \mathbb{E}\left[ \left| \left( \sum_{u \neq v} X_u Y_v S(u) \overline{S(v)} \mathbf{1}_{[H(u)=H(v)]} \right) \right|^2 \right]. \tag{32}$$

Using the fact that functions $S$ and $H$ are independent and Lemma 6 (proved below), we can bound the expectation of the second non-diagonal term in the above equation.

$$\mathbb{E}\left[\left|\left(\sum_{u\neq v} X_u Y_v S(u)\overline{S(v)}\mathbf{1}_{[H(u)=H(v)]}\right)\right|^2\right] = \mathbb{E}\Bigg[\sum_{\substack{u_1\neq v_1\\ u_2\neq v_2}} X_{u_1}Y_{v_1}X_{u_2}Y_{v_2}\times\cdots$$

$$\cdots\times S(u_1)\overline{S(v_1)S(u_2)}S(v_2)\mathbf{1}_{[H(u_1)=H(v_1)]}\mathbf{1}_{[H(u_2)=H(v_2)]}\Bigg], \tag{33}$$

$$= \sum_{\substack{u_1\neq v_1\\ u_2\neq v_2}} \mathbb{E}\Big[X_{u_1}Y_{v_1}X_{u_2}Y_{v_2}S(u_1)\overline{S(v_1)S(u_2)}S(v_2)\Big]\cdot\mathbb{E}[\mathbf{1}_{[H(u_1)=H(v_1)]}\mathbf{1}_{[H(u_2)=H(v_2)]}], \tag{34}$$

$$\leq \frac{1}{D}\sum_{\substack{u_1\neq v_1\\ u_2\neq v_2}} \mathbb{E}\Big[X_{u_1}Y_{v_1}X_{u_2}Y_{v_2}S(u_1)\overline{S(v_1)S(u_2)}S(v_2)\Big], \tag{35}$$

$$\leq \frac{1}{D}\sum_{\substack{u_1\neq v_1\\ u_2\neq v_2}} \mathbb{E}\Big[|X_{u_1}||Y_{v_1}||X_{u_2}||Y_{v_2}|S(u_1)\overline{S(v_1)S(u_2)}S(v_2)\Big], \tag{36}$$

$$= \frac{1}{D}\mathbb{E}\left[\left|\left(\sum_{u\neq v\in[d]^p} |X_u||Y_v|S(u)\overline{S(v)}\right)\right|^2\right]. \tag{37}$$

We bound the above equation using the second-moment bound of Lemma 6. Therefore, we begin by restating the second-moment bound in the proof of Lemma 6,

$$\mathbb{E}\left[\left|\left(\sum_{u,v\in[d]^p} |X_u||Y_v|S(u)\overline{S(v)}\right)\right|^2\right] = \left(\langle\mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \sum_{i=1}^d x_i^2y_i^2\right)^p. \tag{38}$$

Now, we expand the term $\left|\left(\sum_{u,v\in[d]^p}|X_u||Y_v|S(u)\overline{S(v)}\right)\right|^2$ from the above equation as follows,

$$\left|\left(\sum_{u,v\in[d]^p} |X_u||Y_v|S(u)\overline{S(v)}\right)\right|^2 = \left|\sum_{u\in[d]^p}|X_u||Y_v| + \sum_{\substack{u,v\in[d]^p\\ u\neq v}}|X_u||Y_v|S(u)\overline{S(v)}\right|^2, \tag{39}$$

$$= \left(\sum_{u\in[d]^p}|X_u||Y_u| + \sum_{\substack{u,v\in[d]^p\\ u\neq v}}|X_u||Y_v|S(u)\overline{S(v)}\right)\times\cdots$$

$$\cdots\times\overline{\left(\sum_{u\in[d]^p}|X_u||Y_u| + \sum_{\substack{u,v\in[d]^p\\ u\neq v}}|X_u||Y_v|S(u)\overline{S(v)}\right)}, \tag{40}$$

$$= \left(\sum_{u\in[d]^p}|X_u||Y_u| + \sum_{\substack{u,v\in[d]^p\\ u\neq v}}|X_u||Y_v|S(u)\overline{S(v)}\right)\times\cdots$$

$$\cdots\times\left(\sum_{u\in[d]^p}|X_u||Y_u| + \sum_{\substack{u,v\in[d]^p\\ u\neq v}}|X_u||Y_v|\overline{S(u)}S(v)\right). \tag{41}$$

By further expanding the RHS of the above equation, we get

$$\begin{aligned}\left|\left(\sum_{u,v\in[d]^p}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\right|^2 &= \sum_{u_1,u_2\in[d]^p}|X_{u_1}|\,|Y_{u_1}|\,|X_{u_2}|\,|Y_{u_2}|\,+\cdots\\ \cdots+\sum_{u_1\in[d]^p}\sum_{\substack{u_2,v_2\in[d]^p\\ u_2\neq v_2}}&|X_{u_1}|\,|Y_{u_1}|\,|X_{u_2}|\,|Y_{v_2}|\,\overline{S(u_2)}S(v_2)+\cdots\\ \cdots+\sum_{\substack{u_1,v_1\in[d]^p\\ u_1\neq v_1}}\sum_{u_2\in[d]^p}&|X_{u_1}|\,|Y_{v_1}|\,|X_{u_2}|\,|Y_{u_2}|\,S(u_1)\overline{S(v_1)}\,+\cdots\\ \cdots+\sum_{\substack{u_1,v_1\in[d]^p\\ u_1\neq v_1}}\sum_{\substack{u_2,v_2\in[d]^p\\ u_2\neq v_2}}&|X_{u_1}|\,|Y_{v_1}|\,|X_{u_2}|\,|Y_{v_2}|\,S(u_1)\overline{S(v_1)}\,\overline{S(u_2)}S(v_2).\end{aligned}\tag{42}$$

We know that $u_2\neq v_2,\forall u_2,v_2\in[d]^p$,

$$\sum_{u_1\in[d]^p}\sum_{\substack{u_2,v_2\in[d]^p\\ u_2\neq v_2}}|X_{u_1}|\,|Y_{u_1}|\,|X_{u_2}|\,|Y_{v_2}|\,\mathbb{E}[\overline{S(u_2)}S(v_2)]=0,\tag{43}$$

as $\mathbb{E}\Big[S(u_2)\overline{S(v_2)}\Big]=0,\,\forall\,u_2\neq v_2\in[d]^p$. Similarly, for $u_1\neq v_1,\forall u_1,v_1\in[d]^p$,

$$\sum_{\substack{u_1,v_1\in[d]^p\\ u_1\neq v_1}}\sum_{u_2\in[d]^p}|X_{u_1}|\,|Y_{v_1}|\,|X_{u_2}|\,|Y_{u_2}|\,\mathbb{E}[S(u_1)\overline{S(v_1)}]=0,\tag{44}$$

as $\mathbb{E}\Big[\overline{S(u_1)}S(v_1)\Big]=0,\,\forall\,u_1\neq v_1\in[d]^p$. Substituting this into Equation (42) upon computing expectation, we get

$$\begin{aligned}\mathbb{E}\left|\left(\sum_{u,v\in[d]^p}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\right|^2 &= \sum_{u_1,u_2\in[d]^p}|X_{u_1}|\,|Y_{u_1}|\,|X_{u_2}|\,|Y_{u_2}|\,+\cdots\\ &\cdots+\mathbb{E}\Bigg[\sum_{\substack{u_1,v_1\in[d]^p\\ u_1\neq v_1}}\sum_{\substack{u_2,v_2\in[d]^p\\ u_2\neq v_2}}|X_{u_1}|\,|Y_{v_1}|\,|X_{u_2}|\,|Y_{v_2}|\;\times\cdots\\ &\cdots\times S(u_1)\overline{S(v_1)}\,\overline{S(u_2)}S(v_2)\Bigg],\\ &=\langle X,Y\rangle^2+\mathbb{E}\left|\left(\sum_{u\neq v}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\right|^2.\end{aligned}$$

Now, we conclude that,

$$\mathbb{E}\left|\left(\sum_{u\neq v}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\right|^2=\mathbb{E}\left|\left(\sum_{u,v\in[d]^p}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\right|^2-\langle\mathbf{x},\mathbf{y}\rangle^{2p}.\tag{45}$$

Now substitute the value of $\mathbb{E}\left[\left|\left(\sum_{u,v\in[d]^p}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\right|^2\right]$ from Equation (38), we get

$$\mathbb{E}\left|\left(\sum_{u\neq v}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\right|^2=\left(\langle\mathbf{x},\mathbf{y}\rangle^2+\|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2-\sum_{i=1}^{d}x_i^2y_i^2\right)^p-\langle\mathbf{x},\mathbf{y}\rangle^{2p}.\tag{46}$$

Further we substitute this value in Equation (37), we get

$$\mathbb{E}\left[\left|\left(\sum_{u\neq v} X_u Y_v S(u)\overline{S(v)}\mathbf{1}_{[H(u)=H(v)]}\right)\right|^2\right] \leq \frac{1}{D}\left(\left(\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \sum_{i=1}^d x_i^2 y_i^2\right)^p - \langle \mathbf{x},\mathbf{y}\rangle^{2p}\right). \quad (47)$$

Now we can compute the second moment using Equation (32) as follows,

$$\mathbb{E}\left[|\hat{k}_C(\mathbf{x},\mathbf{y})|^2\right] = \langle X, Y\rangle^2 + \mathbb{E}\left[\left|\left(\sum_{u\neq v} X_u Y_v S(u)\overline{S(v)}\mathbf{1}_{[H(u)=H(v)]}\right)\right|^2\right], \quad (48)$$

$$\leq \langle \mathbf{x},\mathbf{y}\rangle^{2p} + \frac{1}{D}\left(\left(\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \sum_{i=1}^d x_i^2 y_i^2\right)^p - \langle \mathbf{x},\mathbf{y}\rangle^{2p}\right). \quad (49)$$

Now, compute variance as follows

$$\mathrm{Var}(\hat{k}_C(\mathbf{x},\mathbf{y})) = \mathbb{E}\left[|\hat{k}_C(\mathbf{x},\mathbf{y})|^2\right] - \left|\mathbb{E}\left[\hat{k}_C(\mathbf{x},\mathbf{y})\right]\right|^2, \quad (50)$$

$$\leq \langle \mathbf{x},\mathbf{y}\rangle^{2p} + \frac{1}{D}\left(\left(\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \sum_{i=1}^d x_i^2 y_i^2\right)^p - \langle \mathbf{x},\mathbf{y}\rangle^{2p}\right) - \langle \mathbf{x},\mathbf{y}\rangle^{2p}, \quad (51)$$

$$= \frac{1}{D}\left(\left(\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \sum_{i=1}^d x_i^2 y_i^2\right)^p - \langle \mathbf{x},\mathbf{y}\rangle^{2p}\right). \quad (52)$$

We can upper bound the above equation by using inequality $\langle \mathbf{x},\mathbf{y}\rangle^2 \leq \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2$, then

$$\mathrm{Var}(\hat{k}_C(\mathbf{x},\mathbf{y})) \leq \frac{1}{D}\left(\left(2\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2\right)^p - \|\mathbf{x}\|_2^{2p}\|\mathbf{y}\|_2^{2p}\right), \quad (53)$$

$$= \frac{(2^p-1)}{D}\|\mathbf{x}\|_2^{2p}\|\mathbf{y}\|_2^{2p}. \quad (54)$$

□

**Remark 4** (Sketching time for Complex `TensorSketch`)**.** *Let $\mathbf{x} \in \mathbb{R}^d$ and let $p \geq 1$ be an integer. The Complex* ***`TensorSketch`*** *of $\mathbf{x}^{\otimes p}$ with sketch dimension $D$ can be computed in $O(p(\mathrm{nnz}(\mathbf{x}) + D\log D))$ time. This follows because* ***`TensorSketch`*** *avoids explicitly forming the tensor $\mathbf{x}^{\otimes p}$. Instead, it applies $p$ independent Complex* ***`CountSketch`*** *to $\mathbf{x}$, each takes $O(\mathrm{nnz}(\mathbf{x}))$ time, and combines the resulting $p$ sketches using circular convolution, which is implemented via FFT in $O(pD\log D)$ time.*

**Remark 5** (Comparison between Real & Complex `TensorSketch`)**.** *For degree-$p$ polynomial kernel estimation, the Complex* ***`TensorSketch`*** *estimator has variance less than or equal to that of the real* ***`TensorSketch`*** *estimator when $\langle \mathbf{x},\mathbf{y}\rangle^2 - \sum_{i=1}^d x_i^2 y_i^2 \geq 0$.*

*Let's analyze this using the variance bounds,*

$$\mathrm{Var}_{TS\text{-}real} \leq \frac{1}{D}\left(\left(2\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - 2\sum_{i=1}^d x_i^2 y_i^2\right)^p - \langle \mathbf{x},\mathbf{y}\rangle^{2p}\right),$$

$$\mathrm{Var}_{TS\text{-}complex} \leq \frac{1}{D}\left(\left(\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \sum_{i=1}^d x_i^2 y_i^2\right)^p - \langle \mathbf{x},\mathbf{y}\rangle^{2p}\right).$$

*Their difference is*

$$\mathrm{Var}_{TS\text{-}real} - \mathrm{Var}_{TS\text{-}complex} \leq \frac{1}{D}\left(2\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - 2\sum_{i=1}^d x_i^2 y_i^2\right)^p \cdots$$

$$\cdots - \frac{1}{D}\left(\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \sum_{i=1}^d x_i^2 y_i^2\right)^p. \quad (55)$$

*Define*

$$A := \langle \mathbf{x}, \mathbf{y} \rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2, \tag{56}$$

$$B := \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2. \tag{57}$$

*Then*

$$\mathrm{Var}_{\textit{TS-real}} - \mathrm{Var}_{\textit{TS-complex}} \leq \frac{1}{D} \left( (2A + B)^p - (A + B)^p \right). \tag{58}$$

*Using the binomial theorem,*

$$(2A + B)^p = \sum_{k=0}^{p} \binom{p}{k} B^{p-k} (2A)^k, \tag{59}$$

$$(A + B)^p = \sum_{k=0}^{p} \binom{p}{k} B^{p-k} A^k. \tag{60}$$

*Subtracting gives*

$$(2A + B)^p - (A + B)^p = \sum_{k=0}^{p} \binom{p}{k} B^{p-k} \left( (2A)^k - A^k \right) \tag{61}$$

$$= \sum_{k=0}^{p} \binom{p}{k} B^{p-k} (2^k - 1) A^k. \tag{62}$$

*Since the $k = 0$ term vanishes ($2^0 - 1 = 0$), we obtain*

$$\mathrm{Var}_{\textit{TS-real}} - \mathrm{Var}_{\textit{TS-complex}} \leq \frac{1}{D} \sum_{k=1 \mathrm{Var}_{\mathrm{real}}}^{p} \binom{p}{k} (2^k - 1) \, A^k B^{p-k} \tag{63}$$

$$= \frac{1}{D} \sum_{k=1}^{p} \binom{p}{k} (2^k - 1) \left( \langle \mathbf{x}, \mathbf{y} \rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2 \right)^k \left( \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \right)^{p-k}. \tag{64}$$

*If $\langle \mathbf{x}, \mathbf{y} \rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2 \geq 0$, Then*

$$\mathrm{Var}_{\textit{TS-real}} - \mathrm{Var}_{\textit{TS-complex}} \geq 0. \tag{65}$$

| Algorithm | Variance | Time Complexity |
|---|---|---|
| `TensorSketch` (Complex) | $\frac{1}{D}\left(\left(\langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^p - \langle \mathbf{x}, \mathbf{y} \rangle^{2p}\right)$ | $O(p\,(\mathrm{nnz}(\mathbf{x}) + \mathrm{nnz}(\mathbf{y}) + D \log D))$ |
| `TensorSketch` (Real) Pham & Pagh (2013) | $\frac{1}{D}\left(\left(2\langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - 2\sum_{i=1}^{d} x_i^2 y_i^2\right)^p - \langle \mathbf{x}, \mathbf{y} \rangle^{2p}\right)$ | $O(p\,(\mathrm{nnz}(\mathbf{x}) + \mathrm{nnz}(\mathbf{y}) + D \log D))$ |
| JL (Complex Rademacher) Wacker et al. (2024) | $\frac{1}{D}\left(\left(\|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 + \langle \mathbf{x}, \mathbf{y} \rangle^2 - \sum_{k=1}^{d} x_k^2 y_k^2\right)^p - \langle \mathbf{x}, \mathbf{y} \rangle^{2p}\right)$ | $O(pDd)$ |
| JL (Complex Gaussian) Wacker et al. (2024) | $\frac{1}{D}\left(\left(\|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 + \langle \mathbf{x}, \mathbf{y} \rangle^2\right)^p - \langle \mathbf{x}, \mathbf{y} \rangle^{2p}\right)$ | $O(pDd)$ |

Table 2: **Comparison of variance bounds and sketching time for complex `TensorSketch` and baseline methods.** Here, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ are input vectors, $p$ denotes the polynomial degree, and $D$ is the sketch dimension. The quantity $\mathrm{nnz}(\mathbf{x})$ represents the number of nonzero entries in $\mathbf{x}$. All listed methods yield unbiased estimators of the inner product between the vectors $\mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p} \in \mathbb{R}^{d^p}$, i.e., $\langle \mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p} \rangle$.

We begin by stating a lemma that serves as a key step to bound the second moment of the estimator in the proof of Theorem 3 .

**Lemma 6.** *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, let $p > 1$ be an integer, and let $s_1, \ldots, s_p : [d] \to \{1, \omega, \omega^2, \omega^3\}$ be independent functions, each taking values uniformly from the four fourth roots of unity. Define*

$$Z \;=\; \prod_{j=1}^{p} Z_{s_j}(\mathbf{x})\,\overline{Z_{s_j}(\mathbf{y})}, \tag{66}$$

*Where*

$$Z_{s_j}(\mathbf{x}) = \sum_{i=1}^{d} x_i\, s_j(i), \qquad\qquad Z_{s_j}(\mathbf{y}) = \sum_{i=1}^{d} y_i\, s_j(i). \tag{67}$$

*Then,*

$$\mathbb{E}[Z] = \langle \mathbf{x}, \mathbf{y} \rangle^p, \tag{68}$$

$$\mathrm{Var}[Z] = \left( \langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2 \right)^p \;-\; \langle \mathbf{x}, \mathbf{y} \rangle^{2p}, \tag{69}$$

$$\leq 2^p \, \|\mathbf{x}\|_2^{2p} \, \|\mathbf{y}\|_2^{2p}. \tag{70}$$

*Proof.* Following the approach of Braverman et al. (2010), adapted from Pham & Pagh (2013)[Lemma 8], we compute the expectation and variance of $Z$. First, we consider the expectation. For each $j$, we note that

$$\mathbb{E}\Big[ Z_{s_j}(\mathbf{x})\,\overline{Z_{s_j}(\mathbf{y})} \Big] = \mathbb{E}\left[ \left( \sum_{i=1}^{d} x_i\, s_j(i) \right) \left( \sum_{k=1}^{d} y_k\, \overline{s_j(k)} \right) \right], \tag{71}$$

$$= \sum_{i=1}^{d} \sum_{k=1}^{d} x_i y_k \, \mathbb{E}[s_j(i)\overline{s_j(k)}], \tag{72}$$

$$= \sum_{i=1}^{d} x_i y_i \, \mathbb{E}[|s_j(i)|^2] + \sum_{i \neq k} x_i y_k \, \mathbb{E}[s_j(i)\overline{s_j(k)}], \tag{73}$$

$$= \langle \mathbf{x}, \mathbf{y} \rangle, \tag{74}$$

Where, $\mathbb{E}[s_j(i)\overline{s_j}(k)] = 0, \forall i \neq k$ and $\mathbb{E}[|s_j(i)|^2] = 1, \forall\ i \in [d]$.

Since the functions $s_j$ are independent across different $j$, we have

$$\mathbb{E}[Z] = \prod_{j=1}^{p} \mathbb{E}[Z_{s_j}(\mathbf{x})\overline{Z_{s_j}(\mathbf{y})}] = \langle \mathbf{x}, \mathbf{y} \rangle^p. \tag{75}$$

Next, to bound the variance,

$$\mathrm{Var}(Z) = \mathbb{E}[|Z|^2] - |(\mathbb{E}[Z])|^2. \tag{76}$$

Because functions is independent across different $j$, we may write

$$\mathbb{E}[|Z|^2] = \prod_{j=1}^{p} \mathbb{E}\Big[ \big| \left( Z_{s_j}(\mathbf{x})\,\overline{Z_{s_j}(\mathbf{y})} \right) \big|^2 \Big]. \tag{77}$$

For each $j$, expanding the square gives

$$\mathbb{E}\Big[ \big| \left( Z_{s_j}(\mathbf{x})\,\overline{Z_{s_j}(\mathbf{y})} \right) \big|^2 \Big] = \mathbb{E}\left[ \left( \sum_{i=1}^{d} x_i\, s_j(i) \right) \left( \sum_{k=1}^{d} y_k\, \overline{s_j(k)} \right) \left( \sum_{i=1}^{d} x_i\, \overline{s_j(i)} \right) \left( \sum_{k=1}^{d} y_k\, s_j(k) \right) \right], \tag{78}$$

$$= \sum_{i=1}^{d} \sum_{i'=1}^{d} \sum_{k=1}^{d} \sum_{k'=1}^{d} x_i\, x_{i'}\, y_k\, y_{k'} \, \mathbb{E}\Big[ s_j(i)\overline{s_j(k) s_j(i')} s_j(k') \Big]. \tag{79}$$

Observing that $\mathbb{E}[s_j(i)\overline{s_j(k)s_j(i')}s_j(k')]$ is nonzero only when the indices form pairs (including the possibility that all four are identical), we have

$$\mathbb{E}[s_j(i)\overline{s_j(k)s_j(i')}s_j(k')] = \begin{cases} 1, & \text{if } i = k = i' = k', \\ 1, & \text{if } i = k \neq i' = k', \\ 1, & \text{if } i = i' \neq k = k', \\ 0, & \text{otherwise.} \end{cases} \tag{80}$$

The contribution from terms with $i = k = i' = k'$ is

$$\sum_{i=1}^{d} x_i^2 y_i^2. \tag{81}$$

Terms with $i = k \neq i' = k'$ contribute

$$\sum_{i \neq i'} x_i y_i \, x_{i'} y_{i'} = \left( \sum_{i=1}^{d} x_i y_i \right)^2 - \sum_{i=1}^{d} x_i^2 y_i^2 = \langle \mathbf{x}, \mathbf{y} \rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2. \tag{82}$$

Finally, for $i = i' \neq k = k'$ we obtain

$$\sum_{i \neq k} x_i^2 y_k^2 = \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2. \tag{83}$$

Thus, summing these contributions, we have

$$\mathbb{E}\big[ | \left( Z_{s_j}(\mathbf{x}) \, Z_{s_j}(\mathbf{y}) \right) |^2 \big] = \sum_{i=1}^{d} x_i^2 y_i^2 + \left( \langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - 2 \sum_{i=1}^{d} x_i^2 y_i^2 \right), \tag{84}$$

$$= \langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2. \tag{85}$$

Substituting this bound into Equation (77) yields

$$\mathbb{E}[|Z|^2] = \left( \langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2 \right)^p, \tag{86}$$

Which completes the proof since

$$\mathrm{Var}(Z) = \mathbb{E}[|Z|^2] - \langle \mathbf{x}, \mathbf{y} \rangle^{2p}, \tag{87}$$

$$= \left( \langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2 \right)^p - \langle \mathbf{x}, \mathbf{y} \rangle^{2p}. \tag{88}$$

Using the Cauchy–Schwarz inequality, $\langle \mathbf{x}, \mathbf{y} \rangle^2 \leq \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2$, and noting that $\sum_{i=1}^{d} x_i^2 y_i^2 \geq 0$, it follows that

$$\mathrm{Var}(Z) \leq 2^p \|\mathbf{x}\|_2^{2p} \|\mathbf{y}\|_2^{2p}. \tag{89}$$

□

### 4.3 Complex `Recursive TensorSketch`

We now define Complex `Recursive TensorSketch` which gives sketching algorithm for $\mathbf{x}^{\otimes p} \in \mathbb{R}^{d^p}$, where $\mathbf{x} \in \mathbb{R}^d$, which consists of a hierarchical binary tree composition of degree-2 complex `TensorSketch`. This method provides an unbiased estimate of the degree-$p$ polynomial kernel and give a variance bounds independent of degree-$p$.

**Definition 4.3: Complex `Recursive TensorSketch` Mapping**

Given a vector $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{x}^{\otimes p} \in \mathbb{R}^{d^p}$ where $p$ takes values that are powers of two, then the Complex `Recursive TensorSketch` is a randomized linear map

$$\Pi^p : \mathbb{R}^{d^p} \to \mathbb{C}^D, \text{defined as,} \;\; \Pi^p := Q^p \cdot T^p, \;\; \text{where}$$

- $T^p = T_1 \otimes T_2 \otimes \cdots \otimes T_p$, with each $T_i \in \mathbb{R}^{D \times d} \quad \forall i \in [p]$ a Complex `CountSketch` matrix (Definition 4.1),
- $Q^p = S^2 \cdot S^4 \cdots S^{p/2} \cdot S^p$, with each $S^\ell \in \mathbb{R}^{D^{\ell/2} \times D^\ell}$ a Kronecker product of matrices $S^\ell_j \in \mathbb{R}^{D \times D^2}$,
- Each $S^\ell_j$ is a Complex `TensorSketch` matrix of degree 2 (Definition 4.2), and $S^\ell = S^\ell_1 \otimes S^\ell_2 \otimes \cdots \otimes S^\ell_{\ell/2}$.

The following theorem establishes that the complex `Recursive TensorSketch` provides an unbiased estimator for the degree-$p$ polynomial kernel, derives variance bound and analyze time complexity.

**Theorem 7** (Complex `Recursive TensorSketch`)**.** *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and let $p$ be a power of two. Let*

$$\Pi^p = Q^p \, T^p : \mathbb{R}^{d^p} \to \mathbb{C}^D \tag{90}$$

*denote the Complex `Recursive TensorSketch` map stated in Definition 4.3. Define the estimator*

$$\hat{k}(\mathbf{x}, \mathbf{y}) \;:=\; \left\langle \Pi^p \mathbf{x}^{\otimes p}, \overline{\Pi^p \mathbf{y}^{\otimes p}} \right\rangle. \tag{91}$$

*Then the estimator $\hat{k}_C(\mathbf{x}, \mathbf{y})$ is unbiased and satisfies*

$$\mathbb{E}\Big[\hat{k}_C(\mathbf{x}, \mathbf{y})\Big] = \langle \mathbf{x}, \mathbf{y} \rangle^p, \tag{92}$$

$$\mathrm{Var}\Big[\hat{k}_C(\mathbf{x}, \mathbf{y})\Big] \le \frac{2}{D} \|\mathbf{x}\|_2^{2p} \|\mathbf{y}\|_2^{2p}. \tag{93}$$

*Moreover, the sketch $\Pi^p \mathbf{x}^{\otimes p}$ can be computed in $O(p(\mathrm{nnz}(\mathbf{x}) + D \log D))$ time.*

We defer the proof to later in the section and begin by stating some intermediate results that will be key ingredients in proving the theorem. The following lemmas establish moment bounds for the Complex `Recursive TensorSketch` and are used to prove Theorem 7.

**Lemma 8.** *Let $\Pi^p = Q^p \cdot T^p : \mathbb{R}^{d^p} \to \mathbb{C}^D$ denote the Complex `Recursive TensorSketch` map stated in Definition 4.3. Then for all $i \in [D]$ and $j \in [d^p]$, the following moments hold:*

$$\mathbb{E}[\Pi^p_{ij}] = 0, \tag{94}$$

$$\mathbb{E}\big[\, |\Pi^p_{ij}|^2 \,\big] = \frac{1}{D}, \tag{95}$$

$$\mathbb{E}\big[(\Pi^p_{ij})^2\big] = 0, \tag{96}$$

$$\mathbb{E}\big[\, |\Pi^p_{ij}|^4 \,\big] = \frac{1}{D} + \frac{2}{D^{2p}} \left( \prod_{i=1}^{p/2} (D^{2i} - 1) \right). \tag{97}$$

*Proof.* The entries of `Recursive TensorSketch` can be constructed as follows

$$\Pi^p = Q^p \cdot T^p \tag{98}$$

By expanding it we can get

$$\begin{aligned}
\Pi^p &= Q^{p/2}\left(S_1^p \otimes S_2^p \otimes \cdots \otimes S_{p/2-1}^p \otimes S_{p/2}^p\right)(T_1 \otimes T_2 \otimes \cdots \otimes T_{p-1} \otimes T_p) \\
&\vdots \\
&= S_1^2\Bigg[S_1^4\Bigg(\cdots S_1^{p/2}\left(S_1^p(T_1 \otimes T_2) \otimes S_2^p(T_3 \otimes T_4)\right)\Bigg) \otimes \cdots \\
&\cdots \otimes S_2^4\Bigg(\cdots S_{p/4}^{p/2}\left(S_{p/2-1}^p(T_{p-3} \otimes T_{p-2}) \otimes S_{p/2}^p(T_{p-1} \otimes T_p)\right)\Bigg)\Bigg]
\end{aligned} \tag{99}$$

We know that each $T_p$ corresponds to a `CountSketch` matrix and each $S_a^\ell$ represents a `TensorSketch` transformation. Consequently, we may express the entry as follows

For each `CountSketch` matrix $T_p$, the $j$-th entry is

$$(T_p)_j = \sum_{i=1}^{d} \sigma_{T_p}(i)\mathbf{1}_{h(i)=j} \tag{100}$$

Where, Let mapping $h : [d] \to [D]$ to map features into $k$ buckets and also let mapping $\sigma_{T_p} : [d] \to \{1, \omega, \omega^2, \omega^3\}$ for each feature. To capture feature contributions, we introduce an indicator random variable $\mathbf{1}_{h(i)=j}$

$$\mathbf{1}_{h(i)=j} = \begin{cases} 1, & \text{if } h(i) = j, \\ 0, & \text{otherwise.} \end{cases}$$

For each `TensorSketch` matrix $S_a^\ell$, the $k$-th entry is

$$(S_a^\ell)_k = \sum_{i=1}^{D}\sum_{j=1}^{D} \phi_{S_a^\ell}(i,j)\Upsilon_{ij}^k \tag{101}$$

Where, Let mapping $h_1, h_2 : [D] \to [D]$ and also let mapping $\phi_1, \phi_2 : [D] \to \{1, \omega, \omega^2, \omega^3\}$, so we can write

- $H(i,j) = [h_1(i) + h_2(j)] \bmod \mathrm{D}$,
- $\phi_{S_a^\ell}(i,j) = \phi_1(i)\overline{\phi_2(j)}$.

We define indicator variable $\Upsilon_{ij}^k$ as:

$$\Upsilon_{ij}^k = \begin{cases} 1, & \text{if } H(i,j) = k, \\ 0, & \text{otherwise.} \end{cases}$$

By substituting the expressions from Equations (100) and (101) into the recursive formulation (99), the $(i,j)$-th entry of $\Pi^p$ can be written explicitly as follows

$$\begin{aligned}
\Pi_{ij}^p = \sum_{\mathbf{l}\in[D]^p}^{D} &\left(\sigma_{T_1}(j_1)\cdots\sigma_{T_p}(j_p)\prod_{k=1}^{p} z_{j_k l_k^p}\right)\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right)\Upsilon_{l_a^{q/2} l_{2a-1}^q l_{2a}^q}^{l_a^{q/2}}\right)\times\cdots \\
&\cdots \times \phi_{S_1^2}(l_1^2, l_2^2)\Upsilon_{l_1^2 l_2^2}^{i},
\end{aligned} \tag{102}$$

where the lexicographic index $j$ is given by

$$j = 1 + \sum_{i=1}^{p} (j_i - 1) d^{i-1},$$

Since the function in both `CountSketch` and `TensorSketch` takes value uniformly from the fourth roots of unity and satifies

$$\mathbb{E}[s(r)] = 0, \qquad \mathbb{E}[\,|s(r)|^2\,] = 1, \qquad \mathbb{E}[\,s(r)^2\,] = 0. \tag{103}$$

**First Moment:**

$$\mathbb{E}\left[\Pi_{ij}^{p}\right] = \mathbb{E}\Bigg[ \sum_{\mathbf{l}\in[D]^p} \left( \sigma_{T_1}(j_1) \cdots \sigma_{T_p}(j_p) \prod_{k=1}^{p} z_{j_k l_k^p} \right) \prod_{q=4,8,\ldots,p} \prod_{a=1}^{q/2} \phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right) \times \cdots$$

$$\cdots \times \Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q} \phi_{S_1^2}(l_1^2, l_2^2) \Upsilon^{i}_{l_1^2 l_2^2} \Bigg], \tag{104}$$

We observe that for every term in the sum, the product $\sigma_{T_1}(j_1) \cdots \sigma_{T_p}(p)$ involves the random function whose values are drawn independently and uniformly from the four fourth roots of unity, therefore $\mathbb{E}[\sigma_{T_i}(i)] = 0$ for each $i \in [p]$ as stated in Equation (103). Therefore, we can write

$$\mathbb{E}\left[\Pi_{ij}^{p}\right] = 0. \tag{105}$$

**Second Moment Norm:**

Let $Z = |\Pi_{ij}^{p}|^2$. We can decompose $Z$ as $Z = Z_1 + Z_2$, where $Z_1$ is the diagonal part and $Z_2$ is the cross (non-diagonal) part, as detailed below:

$$Z = \left|\Pi_{ij}^{p}\right|^2 \tag{106}$$

$$= \Bigg| \sum_{\mathbf{l}\in[D]^p} \left( \sigma_{T_1}(j_1) \cdots \sigma_{T_p}(j_p) \prod_{k=1}^{p} z_{j_k l_k^p} \right) \left( \prod_{q=4,8,\ldots,p} \prod_{a=1}^{q/2} \phi_{S_a^q}(l_{2a-1}^q, l_{2a}^q) \Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q} \right) \times \cdots$$

$$\times \phi_{S_1^2}(l_1^2, l_2^2) \Upsilon^{i}_{l_1^2 l_2^2} \Bigg|^2$$

$$= Z_1 + Z_2. \tag{107}$$

where

$$Z_1 = \sum_{\mathbf{l}\in[D]^p} \left| \sigma_{T_1}(j_1) \cdots \sigma_{T_p}(j_p) \prod_{k=1}^{p} z_{j_k l_k^p} \right|^2 \times \cdots$$

$$\cdots \times \left| \prod_{q=4,8,\ldots,p} \prod_{a=1}^{q/2} \phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right) \Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q} \right|^2 \times \cdots$$

$$\cdots \times \left|\phi_{S_1^2}(l_1^2, l_2^2)\right|^2 \left|\Upsilon^{i}_{l_1^2 l_2^2}\right|^2, \tag{108}$$

$$Z_2 = \left(\sigma_{T_1}(j_1) \cdots \sigma_{T_p}(j_p)\right)^2 \sum_{\mathbf{l}\neq\mathbf{l}'} \left( \prod_{k=1}^{p} z_{j_k l_k^p} \right) \left( \prod_{k=1}^{p} z_{j_k l_k'^p} \right) \times \cdots$$

$$\cdots \times \prod_{q=4,8,\ldots,p} \prod_{a=1}^{q/2} \phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right) \Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q} \times \cdots$$

$$\cdots \times \prod_{q=4,8,\ldots,p} \prod_{a=1}^{q/2} \phi_{S_a^q}\left(l_{2a-1}'^q, l_{2a}'^q\right) \Upsilon^{l_a'^{q/2}}_{l_{2a-1}'^q l_{2a}'^q} \times \cdots$$

$$\cdots \times \phi_{S_1^2}(l_1^2, l_2^2)\, \phi_{S_1^2}(l_1'^2, l_2'^2)\, \Upsilon^{i}_{l_1^2 l_2^2}\, \Upsilon^{i}_{l_1'^2 l_2'^2}. \tag{109}$$

Applying expectation, we have,

$$\mathbb{E}[Z_1] = \mathbb{E}\left[\sum_{\mathbf{l}\in[D]^p}\left(\prod_{k=1}^{p}(z_{j_k l_k^p})^2\right)\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}(\Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q})^2\right)(\Upsilon^i_{l_1^2 l_2^2})^2\right], \tag{110}$$

$$= \sum_{\mathbf{l}\in[D]^p}\frac{1}{D^p}\times\frac{1}{D^{p-2}}\times\frac{1}{D}, \tag{111}$$

$$= \frac{1}{D}. \tag{112}$$

and

$$\mathbb{E}[Z_2] = \mathbb{E}\Bigg[\sigma_{T_1}(j_1)\cdots\sigma_{T_p}(j_p)\sum_{\mathbf{l}\neq\mathbf{l}'}\left(\prod_{k=1}^{p} z_{j_k l_k^p}\right)\left(\prod_{k=1}^{p} z_{j_k l_k'^p}\right)\times\cdots \tag{113}$$

$$\cdots\times\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right)\Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}\right)\times\cdots$$

$$\cdots\times\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\phi_{S_a^q}\left(l_{2a-1}'^q, l_{2a}'^q\right)\Upsilon^{l_a'^{q/2}}_{l_{2a-1}'^q l_{2a}'^q}\right)\phi_{S_1^2}(l_1^2, l_2^2)\phi_{S_1^2}(l_1'^2, l_2'^2)(\Upsilon^i_{l_1^2 l_2^2})(\Upsilon^i_{l_1'^2 l_2'^2})\Bigg], \tag{114}$$

$$= 0 \qquad (\because\ \mathbb{E}[\sigma_{T_i}(j_i)] = 0 \quad \forall i\in[p])\,. \tag{115}$$

Therefore, using Equation (112) for $Z_1$ and (115) for $Z_2$,

$$\mathbb{E}\left[\left|\Pi_{ij}^p\right|^2\right] = \mathbb{E}[Z] = \mathbb{E}[Z_1] + \mathbb{E}[Z_2] = \frac{1}{D}. \tag{116}$$

**Second moment $\mathbb{E}[(\Pi_{ij}^p)^2]$.**
Let $Z = \left(\Pi_{ij}^p\right)^2$. We can decompose $Z$ as $Z = Z_1 + Z_2$, where $Z_1$ is the diagonal part and $Z_2$ is the cross (non-diagonal) part, as detailed below:

$$Z = \left(\Pi_{ij}^p\right)^2, \tag{117}$$

$$= \Bigg(\sum_{\mathbf{l}\in[D]^p}\left(\sigma_{T_1}(j_1)\cdots\sigma_{T_p}(j_p)\prod_{k=1}^{p} z_{j_k l_k^p}\right)\times\cdots$$

$$\cdots\times\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\phi_{S_a^q}(l_{2a-1}^q, l_{2a}^q)\Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}\right)\phi_{S_1^2}(l_1^2, l_2^2)\Upsilon^i_{l_1^2 l_2^2}\Bigg)^2, \tag{118}$$

$$= Z_1 + Z_2, \tag{119}$$

where

$$Z_1 = \sum_{\mathbf{l}\in[D]^p}\left(\sigma_{T_1}(j_1)\cdots\sigma_{T_p}(j_p)\prod_{k=1}^{p} z_{j_k l_k^p}\right)^2\times\cdots$$

$$\cdots\times\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right)\Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}\right)^2\phi_{S_1^2}(l_1^2, l_2^2)^2(\Upsilon^i_{l_1^2 l_2^2})^2, \tag{120}$$

$$\tag{121}$$

$$Z_2 = \sigma_{T_1}(j_1)\cdots\sigma_{T_p}(j_p)\sum_{\mathbf{l}\neq\mathbf{l}'}\left(\prod_{k=1}^{p} z_{j_k l_k^p}\right)\left(\prod_{k=1}^{p} z_{j_k l_k'^p}\right)\times\cdots$$
$$\cdots\times\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right)\Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}\right)\times\cdots$$
$$\cdots\times\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\phi_{S_a^q}\left(l_{2a-1}'^q, l_{2a}'^q\right)\Upsilon^{l_a'^{q/2}}_{l_{2a-1}'^q l_{2a}'^q}\right)\phi_{S_1^2}(l_1^2, l_2^2)\phi_{S_1^2}(l_1'^2, l_2'^2)(\Upsilon^i_{l_1^2 l_2^2})(\Upsilon^i_{l_1'^2 l_2'^2}). \tag{122}$$

Therefore, using Equation (103) for $Z_1$ and $Z_2$ we get,

$$\mathbb{E}[Z_1] = 0 \text{ and } \mathbb{E}[Z_2] = 0. \tag{123}$$

Hence,

$$\mathbb{E}\left[\left(\Pi_{ij}^p\right)^2\right] = 0. \tag{124}$$

**Fourth Moment:**

Calculating expectation of 4th moment of $\Pi_{ij}^p$,

$$\left|\Pi_{ij}^p\right|^4 = \left|\sum_{\mathbf{l}\in[D]^p}\left(\sigma_{T_1}(j_1)\cdots\sigma_{T_p}(j_p)\prod_{k=1}^{p} z_{j_k l_k^p}\right)\times\cdots\right.$$
$$\left.\cdots\times\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right)\Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}\right)\phi_{S_1^2}(l_1^2, l_2^2)\Upsilon^i_{l_1^2 l_2^2}\right|^4, \tag{125}$$
$$= \left|\sum_{\mathbf{l}\in[D]^p}\left(\prod_{k=1}^{p} z_{j_k l_k^p}\right)\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right)\Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}\right)\phi_{S_1^2}(l_1^2, l_2^2)\Upsilon^i_{l_1^2 l_2^2}\right|^4,$$

We know that,

$$\left|\sum_n X_n\right|^4 = \left(\sum_n X_n\right)^2\left(\sum_n \overline{X_n}\right)^2,$$
$$= \sum_n |X_n|^4 + 2\sum_{n\neq n_1}|X_n|^2 X_n\overline{X_{n_1}} + 2\sum_{n\neq n_1}|X_n|^2\overline{X_n}X_{n_1} + \cdots$$
$$\cdots + \sum_{n\neq D} X_n^2\overline{X_D}^2 + 2\sum_{n_1\neq n_2}|X_{n_1}|^2|X_{n_2}|^2 + \cdots \tag{126}$$
$$\cdots + 4\sum_{n\neq n_1\neq n_2}|X_n|^2 X_{n_1}\overline{X_{n_2}} + \sum_{n\neq n_1\neq n_2} X_n^2\overline{X_{n_1}X_{n_2}} + \cdots$$
$$\cdots + \sum_{n\neq n_1\neq n_2}\overline{X_n}^2 X_{n_1}X_{n_2} + \sum_{n\neq n_1\neq n_2\neq n_3} X_n X_{n_1}\overline{X_{n_2}X_{n_3}}.$$

Here, we have

$$X_n = \sum_{\mathbf{l}\in[D]^p} \left(\prod_{k=1}^{p} z_{j_k l_k^p}\right)\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2} \phi_{S_a^q}(l_{2a-1}^q, l_{2a}^q)\, \Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}\right)\phi_{S_1^2}(l_1^2, l_2^2)\Upsilon^i_{l_1^2 l_2^2}, \tag{127}$$

$$X_{n_1} = \sum_{\mathbf{l}'\in[D]^p} \left(\prod_{k=1}^{p} z_{j_k l_k'^p}\right)\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2} \phi_{S_a^q}(l_{2a-1}'^q, l_{2a}'^q)\, \Upsilon^{l_a'^{q/2}}_{l_{2a-1}'^q l_{2a}'^q}\right)\phi_{S_1^2}(l_1'^2, l_2'^2)\Upsilon^i_{l_1'^2 l_2'^2}, \tag{128}$$

$$X_{n_2} = \sum_{\mathbf{l}''\in[D]^p} \left(\prod_{k=1}^{p} z_{j_k l_k''^p}\right)\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2} \phi_{S_a^q}(l_{2a-1}''^q, l_{2a}''^q)\, \Upsilon^{l_a''^{q/2}}_{l_{2a-1}''^q l_{2a}''^q}\right)\phi_{S_1^2}(l_1''^2, l_2''^2)\Upsilon^i_{l_1''^2 l_2''^2}, \tag{129}$$

$$X_{n_3} = \sum_{\mathbf{l}'''\in[D]^p} \left(\prod_{k=1}^{p} z_{j_k l_k'''^p}\right)\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2} \phi_{S_a^q}(l_{2a-1}'''^q, l_{2a}'''^q) \times \cdots\right. \tag{130}$$

$$\left. \cdots \times \Upsilon^{l_a'''^{q/2}}_{l_{2a-1}'''^q l_{2a}'''^q}\right)\phi_{S_1^2}(l_1'''^2, l_2'''^2)\Upsilon^i_{l_1'''^2 l_2'''^2},$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[|\Pi_{ij}^p|^4\right] = \mathbb{E}\Bigg[&\sum_n |X_n|^4 + 2\sum_{n\neq n_1} |X_n|^2 X_n \overline{X_{n_1}} + 2\sum_{n\neq n_1} |X_n|^2 \overline{X_n}\, X_{n_1} + \cdots \\ &\cdots + \sum_{n_1\neq n_2} X_{n_1}^2\, \overline{X_{n_2}}^2 + 2\sum_{n_1\neq n_2} |X_{n_1}|^2\, |X_{n_2}|^2 + \cdots \\ &\cdots + 4\sum_{\substack{n\neq n_1\\ n\neq n_2\\ n_1\neq n_2}} |X_n|^2\, X_{n_1}\, \overline{X_{n_2}} + \sum_{\substack{n\neq n_1\\ n\neq n_2\\ n_1\neq n_2}} X_n^2\, \overline{X_{n_1}}\, \overline{X_{n_2}} + \cdots \\ &\cdots + \sum_{\substack{n\neq n_1\\ n\neq n_2\\ n_1\neq n_2}} \overline{X_n}^2\, X_{n_1}\, X_{n_2} + \sum_{n\neq n_1\neq n_2\neq n_3} X_n X_{n_1}\, \overline{X_{n_2}}\, \overline{X_{n_3}}\Bigg].\end{aligned} \tag{131}$$

Odd power $X_n$ has expectation zero because $\mathbb{E}[\sigma_{T_i}(j_i)] = 0$, then we get

$$= \mathbb{E}\left[\sum_n |X_n|^4\right] + \mathbb{E}\left[\sum_{n_1\neq n_2} X_{n_1}^2\, \overline{X_{n_2}}^2\right] + 2\mathbb{E}\left[\sum_{n_1\neq n_2} |X_{n_1}|^2 |X_{n_2}|^2\right]. \tag{132}$$

We analyze each term separately as follows:

$$\begin{aligned}\mathbb{E}\left[\sum_n X_n^4\right] &= \mathbb{E}\Bigg[\Bigg|\sum_{\mathbf{l}\in[D]^p}\left(\prod_{k=1}^{p} z_{j_k l_k^p}\right) \times \cdots \\ &\qquad \cdots \times \left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2} \phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right) \Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}\right) \phi_{S_1^2}(l_1^2, l_2^2)\Upsilon^i_{l_1^2 l_2^2}\Bigg|^4\Bigg],\end{aligned} \tag{133}$$

$$\begin{aligned}&= \mathbb{E}\Bigg[\sum_{\mathbf{l}\in[D]^p}\left(\prod_{k=1}^{p} (z_{j_k l_k^p})^4\right)\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2} \times \cdots\right. \\ &\qquad \left.\cdots \times \left|\phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right)\right|^4 (\Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q})^4\right) \left|\phi_{S_1^2}(l_1^2, l_2^2)\right|^4 (\Upsilon^i_{l_1^2 l_2^2})^4\Bigg],\end{aligned} \tag{134}$$

$$\tag{135}$$

$$= \sum_{\mathbf{l}\in[D]^p} \left(\prod_{k=1}^{p} \mathbb{E}[z_{j_k l_k^p}]\right) \left(\prod_{q=4,8,\ldots,p} \prod_{a=1}^{q/2} \mathbb{E}[\Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}]\right) \mathbb{E}[\Upsilon^i_{l_1^2 l_2^2}], \tag{136}$$

$$= \sum_{\mathbf{l}\in[D]^p} \frac{1}{D^p} \times \frac{1}{D^{p-2}} \times \frac{1}{D}, \tag{137}$$

$$= \frac{1}{D}. \tag{138}$$

$$\begin{aligned}
\mathbb{E}\left[\sum_{n_1\neq n_2} X_{n_1}^2 \overline{X_{n_2}}^2\right] &= \mathbb{E}\Bigg[\sum_{\mathbf{l}\neq\mathbf{l}'} \left(\prod_{k=1}^{p} z_{j_k l_k^p}\right)^2 \left(\prod_{k=1}^{p} z_{j_k l_k'^p}\right)^2 \times \cdots \\
&\quad \cdots \times \left(\prod_{q=4,8,\ldots,p} \prod_{a=1}^{q/2} \phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right) \Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}\right)^2 \times \cdots \\
&\quad \cdots \times \overline{\left(\prod_{q=4,8,\ldots,p} \prod_{a=1}^{q/2} \phi_{S_a^q}\left(l_{2a-1}'^q, l_{2a}'^q\right) \Upsilon^{l_a'^{q/2}}_{l_{2a-1}'^q l_{2a}'^q}\right)^2} \times \cdots \\
&\quad \cdots \times \left(\phi_{S_1^2}(l_1^2, l_2^2)\right)^2 \overline{\left(\phi_{S_1^2}(l_1'^2, l_2'^2)\right)^2} (\Upsilon^i_{l_1^2 l_2^2})^2 (\Upsilon^i_{l_1'^2 l_2'^2})^2\Bigg], \\
&= \mathbb{E}\Bigg[\sum_{\mathbf{l}\neq\mathbf{l}'} \left(\prod_{k=1}^{p} z_{j_k l_k^p}\right)^2 \left(\prod_{k=1}^{p} z_{j_k l_k'^p}\right)^2 \times \cdots \\
&\quad \cdots \times \left(\prod_{q=4,8,\ldots,p} \prod_{a=1}^{q/2} \phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right) \Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}\right)^2 \times \cdots \\
&\quad \cdots \times \overline{\left(\prod_{q=4,8,\ldots,p} \prod_{a=1}^{q/2} \phi_{S_a^q}\left(l_{2a-1}'^q, l_{2a}'^q\right) \Upsilon^{l_a'^{q/2}}_{l_{2a-1}'^q l_{2a}'^q}\right)^2}\Bigg] \times \cdots \\
&\quad \cdots \times \mathbb{E}\left[\left(\phi_{S_1^2}(l_1^2, l_2^2)\right)^2 \overline{\left(\phi_{S_1^2}(l_1'^2, l_2'^2)\right)^2}\right] \mathbb{E}[(\Upsilon^i_{l_1^2 l_2^2})^2] \mathbb{E}[(\Upsilon^i_{l_1'^2 l_2'^2})^2].
\end{aligned} \tag{139}$$

This term vanishes as $\mathbb{E}\left[\left(\phi_{S_1^2}(l_1^2, l_2^2)\right)^2\right] = \left[\overline{\left(\phi_{S_1^2}(l_1'^2, l_2'^2)\right)^2}\right] = 0,$
So,

$$\mathbb{E}\left[\sum_{n_1\neq n_2} X_{n_1}^2 \overline{X_{n_2}}^2\right] = 0. \tag{140}$$

$$2\mathbb{E}\left[\sum_{n_1\neq n_2}|X_{n_1}|^2|X_{n_2}|^2\right] = 2\mathbb{E}\Bigg[\sum_{\mathbf{l}\neq\mathbf{l}'}\left(\prod_{k=1}^{p} z_{j_k l_k^p}\right)^2\left(\prod_{k=1}^{p} z_{j_k l_k'^p}\right)^2 \times \cdots$$
$$\cdots \times \left|\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\phi_{S_a^q}\left(l_{2a-1}^q, l_{2a}^q\right)\Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}\right|^2 \times \cdots$$
$$\cdots \times \left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\phi_{S_a^q}\left(l_{2a-1}'^q, l_{2a}'^q\right)\Upsilon^{l_a'^{q/2}}_{l_{2a-1}'^q l_{2a}'^q}\right)^2 \times \cdots$$
$$\cdots \times \left(\phi_{S_1^2}(l_1^2, l_2^2)\right)^2\left(\phi_{S_1^2}(l_1'^2, l_2'^2)\right)^2(\Upsilon^i_{l_1^2 l_2^2})^2(\Upsilon^i_{l_1'^2 l_2'^2})^2\Bigg]. \tag{141}$$

$$= 2\mathbb{E}\Bigg[\sum_{\mathbf{l}\neq\mathbf{l}'}\left(\prod_{k=1}^{p}(z_{j_k l_k^p})^2\right)\left(\prod_{k=1}^{p}(z_{j_k l_k'^p})^2\right)\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}(\Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q})^2\right)\times\cdots$$
$$\cdots\times\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}(\Upsilon^{l_a'^{q/2}}_{l_{2a-1}'^q l_{2a}'^q})^2\right)(\Upsilon^i_{l_1^2 l_2^2})^2(\Upsilon^i_{l_1'^2 l_2'^2})^2\Bigg], \tag{142}$$
$$= 2\sum_{\mathbf{l}\neq\mathbf{l}'}\left(\prod_{k=1}^{p}\mathbb{E}[z_{j_k l_k^p}]\right)\left(\prod_{k=1}^{p}\mathbb{E}[z_{j_k l_k'^p}]\right)\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\mathbb{E}[\Upsilon^{l_a^{q/2}}_{l_{2a-1}^q l_{2a}^q}]\right)\times\cdots$$
$$\cdots\times\left(\prod_{q=4,8,\ldots,p}\prod_{a=1}^{q/2}\mathbb{E}[\Upsilon^{l_a'^{q/2}}_{l_{2a-1}'^q l_{2a}'^q}]\right)\mathbb{E}[\Upsilon^i_{l_1^2 l_2^2}]\mathbb{E}[\Upsilon^i_{l_1'^2 l_2'^2}], \tag{143}$$
$$= 2\sum_{\mathbf{l}\neq\mathbf{l}'}\frac{1}{D^{2p}}\times\frac{1}{D^{2(p-1)}}, \tag{144}$$
$$= 2\left(\prod_{i=1}^{p/2}D^{2i}(D^{2i}-1)\right)\times\frac{1}{D^{4p-2}}, \tag{145}$$
$$= \frac{2}{D^{2p}}\left(\prod_{i=1}^{p/2}(D^{2i}-1)\right). \tag{146}$$

Therefore,

$$\mathbb{E}\left[(\Pi_{ij}^p)^4\right] = \frac{1}{D} + \frac{2}{D^{2p}}\left(\prod_{i=1}^{p/2}(D^{2i}-1)\right). \tag{147}$$

□

**Proof of Theorem 7**

*Proof.* We first outline the structure of the proof. We assume throughout this proof that the degree $p$ is a power of two. The analysis begins by expressing the estimator as a quadratic form in the sketching matrix, then we prove unbiased estimation by taking expectation. For variance, we computed second moment by decomposing it into diagonal and non-diagonal contributions with respect to the sketching dimension. Using the independence of the functions, we identify all index matchings that give nonzero contribution. The diagonal terms and the non-diagonal terms simplify are computed using moment bound of Complex `Recursive TensorSketch`. Combining these contributions, we obtain a closed-form expression for the variance, which is subsequently upper bounded using the Cauchy-Schwarz inequality, resulting in a variance bound of order $1/D$.

We now present the detailed proof. Let $\Pi^p \in \mathbb{C}^{D\times d^p}$ be Complex `Recursive TensorSketch` , and let $X := \mathbf{x}^{\otimes p}, Y = \mathbf{y}^{\otimes p} \in \mathbb{R}^{d^p}$. Then we $\hat{k}_C(\mathbf{x},\mathbf{y})$ as

$$\hat{k}_C(\mathbf{x},\mathbf{y}) := \langle \Pi^p X, \overline{\Pi^p Y}\rangle. \tag{148}$$

Now, we expand $\hat{k}_C(\mathbf{x},\mathbf{y})$ as follows,

$$\hat{k}_C(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{D}\left(\sum_{j=1}^{d^p}\Pi^p_{ij}X_j\right)\left(\sum_{k=1}^{d^p}\overline{\Pi^p_{ik}}Y_k\right), \tag{149}$$

$$= \sum_{i=1}^{D}\sum_{j,k=1}^{d^p}\Pi^p_{ij}\overline{\Pi^p_{ik}}X_jY_k. \tag{150}$$

Taking the expectation on the both sides, we get

$$\mathbb{E}\left[\hat{k}_C(\mathbf{x},\mathbf{y})\right] = \sum_{i=1}^{D}\sum_{j=1}^{d^p}\mathbb{E}\left[|\Pi^p_{ij}|^2\right]X_jY_j + \sum_{i=1}^{D}\sum_{j\neq k}^{d^p}\mathbb{E}[\Pi^p_{ij}]\mathbb{E}[\overline{\Pi^p_{ik}}]X_jY_k, \tag{151}$$

Using Lemma 8 we have $\mathbb{E}\left[\Pi^p_{ij}\right] = 0$, $\mathbb{E}\left[|\Pi^p_{ij}|^2\right] = \frac{1}{D}$ respectively,

$$\mathbb{E}\left[\hat{k}_C(\mathbf{x},\mathbf{y})\right] = \sum_{i=1}^{D}\frac{1}{D}\sum_{j=1}^{d^p}X_jY_j, \tag{152}$$

$$= \sum_{j=1}^{d^p}X_jY_j, \tag{153}$$

$$= \langle X, Y\rangle = \langle \mathbf{x},\mathbf{y}\rangle^p. \tag{154}$$

The above equation proves that the estimator is unbiased. Now, we compute variance as,

$$\mathrm{Var}\left[\hat{k}_C(\mathbf{x},\mathbf{y})]\right] = \mathbb{E}[|\hat{k}_C(\mathbf{x},\mathbf{y})|^2] - |\mathbb{E}[\hat{k}_C(\mathbf{x},\mathbf{y})]|^2. \tag{155}$$

Thus,

$$\mathbb{E}[|\hat{k}_C(\mathbf{x},\mathbf{y})|^2] = \sum_{i,i'=1}^{D}\sum_{j,k,j',k'=1}^{d^p}\mathbb{E}[\Pi^p_{ij}\overline{\Pi^p_{ik}\Pi^p_{i'j'}}\Pi^p_{i'k'}]X_jY_kX_{j'}Y_{k'}. \tag{156}$$

The non-diagonal term of $\mathbb{E}[|\hat{k}_C(\mathbf{x},\mathbf{y})|^2]$ from Equation (156), corresponding to $i \neq i'$, is non-zero when $j = k$, $j' = k'$, and $j \neq j'$, then,

$$\mathbb{E}[\Pi^p_{ij}\overline{\Pi^p_{ik}\Pi^p_{i'j'}}\Pi^p_{i'k'}] = \mathbb{E}[|\Pi^p_{ij}|^2]\cdot\mathbb{E}[|\Pi^p_{i'j'}|^2], \tag{157}$$

$$= \frac{1}{D^2}\ . \tag{158}$$

So, contribution of non-diagonal part is given by,

$$\sum_{i\neq i'}^{D}\sum_{j=k,j'=k',j\neq j'}\mathbb{E}[\Pi_{ij}^{p}\overline{\Pi_{ik}^{p}\Pi_{i'j'}^{p}}\Pi_{i'k'}^{p}]X_jY_kX_{j'}Y_{k'} = \frac{1}{D^2}\sum_{i\neq i'}\sum_{j\neq j'}X_jY_jX_{j'}Y_{j'}, \tag{159}$$

$$= \frac{D(D-1)}{D^2}\sum_{j\neq j'}X_jY_jX_{j'}Y_{j'}\ . \tag{160}$$

Now for diagonal term of $\mathbb{E}[|\hat{k}_C(\mathbf{x},\mathbf{y})|^2]$ from Equation (156), corresponding to $i=i'$ is given by,

$$\sum_{i=1}^{D}\sum_{j,k,j',k'=1}^{d^p}\mathbb{E}[\Pi_{ij}^{p}\overline{\Pi_{ik}^{p}\Pi_{ij'}^{p}}\Pi_{ik'}^{p}]X_jY_kX_{j'}Y_{k'}. \tag{161}$$

Non-zero contributions occur when

$$\begin{aligned}
&\text{(i) } j=k=j'=k': && \mathbb{E}[|\Pi_{ij}^{p}|^4]X_jY_jX_jY_j,\\
&\text{(ii) } j=j',k=k',j\neq k: && \mathbb{E}[|\Pi_{ij}^{p}|^2]\mathbb{E}[|\Pi_{ik}^{p}|^2]X_jY_jX_kY_k,\\
&\text{(iii) } j=k,j'=k',j\neq j': && \mathbb{E}[|\Pi_{ij}^{p}|^2]\mathbb{E}[|\Pi_{ij'}^{p}|^2]X_jY_jX_{j'}Y_{j'}.
\end{aligned}$$

By using Lemma 8 we have contributions by diagonal term of $\mathbb{E}[|\hat{k}_C(\mathbf{x},\mathbf{y})|^2]$ as,

$$\sum_{i=1}^{D}\left[\left[\frac{1}{D}+\frac{2}{D^{2p}}\left(\prod_{i=1}^{p/2}(D^{2i}-1)\right)\right]\sum_{j=1}^{d^p}X_j^2Y_j^2+\frac{1}{D^2}\left[\sum_{j\neq j'}(X_jY_jX_{j'}Y_{j'}+X_j^2Y_{j'}^2)\right]\right], \tag{162}$$

$$= \left[1+\frac{2}{D^{2p-1}}\left(\prod_{i=1}^{p/2}(D^{2i}-1)\right)\right]\sum_{j=1}^{d^p}X_j^2Y_j^2+\frac{1}{D}\left[\sum_{j\neq j'}(X_jY_jX_{j'}Y_{j'}+X_j^2Y_{j'}^2)\right]. \tag{163}$$

By combining the diagonal and non-diagonal terms using Equation (163) and Equation (160), respectively, we obtain

$$\mathbb{E}[|Z|^2] = \langle X,Y\rangle^2+\left[\frac{2}{D^{2p-1}}\left(\prod_{i=1}^{p/2}(D^{2i}-1)\right)\right]\sum_{j=1}^{d^p}X_j^2Y_j^2 + \frac{1}{D}\left[\sum_{j\neq j'}X_j^2Y_{j'}^2\right]. \tag{164}$$

Now by using Equation (154), we get

$$|\mathbb{E}[\hat{k}_C(\mathbf{x},\mathbf{y})]|^2 = |\mathbb{E}[\langle\Pi^pX,\overline{\Pi^pY}\rangle]|^2, \tag{165}$$

$$= \langle X,Y\rangle^2\ . \tag{166}$$

Now substitute the expressions for $\mathbb{E}[\hat{k}_C(\mathbf{x},\mathbf{y})Z|^2]$ and $|\mathbb{E}[\hat{k}_C(\mathbf{x},\mathbf{y})]|^2$ using Equation (164) and Equation (166) respectively in Equation (155), and we get

$$\text{Var}\left[\hat{k}_C(\mathbf{x},\mathbf{y})\right] = \mathbb{E}[|\hat{k}_C(\mathbf{x},\mathbf{y})|^2]-|\mathbb{E}[\hat{k}_C(\mathbf{x},\mathbf{y})]|^2, \tag{167}$$

$$= \left[\frac{2}{D^{2p-1}}\left(\prod_{i=1}^{p/2}(D^{2i}-1)\right)-\frac{1}{D}\right]\sum_{j=1}^{d^p}X_j^2Y_j^2+\frac{1}{D}\left[\langle X,Y\rangle^2\right], \tag{168}$$

$$\leq \left[\frac{2}{D^{2p-1}}\left(\prod_{i=1}^{p/2}(D^{2i})\right)-\frac{1}{D}\right]\sum_{j=1}^{d^p}X_j^2Y_j^2+\frac{1}{D}\left[\langle X,Y\rangle^2\right]. \tag{169}$$

Now, we upper bound $\prod_{i=1}^{p/2}\left((D)^{2i}-1\right) < \prod_{i=1}^{p/2}\left(D^{2i}\right)$ and further we can simplify the term $\prod_{i=1}^{p/2}\left(D^{2i}\right)$ as,

$$\begin{aligned}\prod_{i=1}^{p/2}\left(D^{2i}\right) =& D^{2\sum_{i=1}^{p/2} i},\\ =& D^{2(1+2+4+\cdots+p/2)},\\ =& D^{2(p-1)}.\end{aligned}$$

Now, substitute this upper bound in Equation 169, we get

$$\leq \left(\frac{2}{D}-\frac{1}{D}\right)\sum_{j=1}^{d^p} X_j^2 Y_j^2 + \frac{1}{D}\left[\langle X, Y\rangle\right], \tag{170}$$

$$= \frac{1}{D}\left[\sum_{j=1}^{d^p} X_j^2 Y_j^2 + \langle X, Y\rangle\right]. \tag{171}$$

Finally, applying the Cauchy–Schwarz inequality,

$$\sum_{j=1}^{d^p} X_j^2 Y_j^2 \leq \|X\|_2^2 \|X\|_2^2, \qquad \langle X, Y\rangle^2 \leq \|X\|_2^2 \|Y\|_2^2, \tag{172}$$

We obtain the variance bound,

$$\operatorname{Var}\left[\langle \Pi^p X, \overline{\Pi^p Y}\rangle\right] \leq \frac{2}{D}\|X\|_2^2\|Y\|_2^2, \tag{173}$$

$$= \frac{2}{D}\|\mathbf{x}\|_2^{2p}\|\mathbf{y}\|_2^{2p} \tag{174}$$

□

**Remark 9.** *The Complex* ***Recursive TensorSketch*** *of $\mathbf{x}^{\otimes p}$ with sketch dimension $D$ can be computed in $O(p(\mathrm{nnz}(\mathbf{x}) + D\log D))$ time. This follows because Complex* ***Recursive TensorSketch*** *avoids explicitly forming the tensor $\mathbf{x}^{\otimes p}$. Instead, it applies $p$ independent Complex* ***CountSketch*** *to $\mathbf{x}$, each takes $O(\mathrm{nnz}(\mathbf{x}))$ time, and then recursively combines them using Complex* ***TensorSketch****. Since the recursion involves $(p-1)$ Complex* ***TensorSketch****, each takes $O(\mathrm{nnz}(\mathbf{x}) + D\log D)$ time, the overall sketching time becomes $O(p(\mathrm{nnz}(\mathbf{x}) + D\log D))$.*

**Corollary 10** (Complex `Recursive TensorSketch` for general degree $p$)**.** *The guarantees of Theorem 7 continue to hold for arbitrary polynomial degrees $p$ that are not powers of two. In this case, we reduce to the nearest larger power of two $p' = 2^{\lceil \log_2 p\rceil}$ and apply the Complex* ***Recursive TensorSketch*** *construction at degree $p'$ after padding the tensor representation with degree $(p'-p)$ first standard basis vector.*

*The resulting estimator therefore remains unbiased for $\langle \mathbf{x}, \mathbf{y}\rangle^p$, satisfies the same variance bound, and has the same asymptotic sketching time as in the power-of-two case.*

*Proof.* We are Following the approach used for this case in the work of Ahle et al. (2020).
Let $p \geq 2$ be arbitrary and set $p' := 2^{\lceil \log_2 p\rceil}$. Instead of working directly with $\mathbf{x}^{\otimes p}$ and $\mathbf{y}^{\otimes p}$, we embed them into order–$p'$ tensors by defining

$$\widetilde{\mathbf{x}}^{\otimes p'} := \mathbf{x}^{\otimes p} \otimes \mathbf{e}_1^{\otimes (p'-p)}, \qquad \widetilde{\mathbf{y}}^{\otimes p'} := \mathbf{y}^{\otimes p} \otimes \mathbf{e}_1^{\otimes (p'-p)},$$

where $\mathbf{e}_1 \in \mathbb{R}^d$ is the first standard basis vector.
We then apply the power-of-two Complex `Recursive TensorSketch` $\Pi^{p'}$ to these padded tensors. Since $\langle \mathbf{e}_1, \mathbf{e}_1\rangle = 1$, this embedding preserves inner products:

$$\left\langle \widetilde{\mathbf{x}}^{\otimes p'}, \widetilde{\mathbf{y}}^{\otimes p'}\right\rangle = \langle \mathbf{x}, \mathbf{y}\rangle^p.$$

Thus, the guarantees established for the power-of-two case transfer directly to general $p$ through this embedding.

□

### 4.4 Concentration Analysis

In this subsection, we analyze the concentration properties of the Complex `CountSketch`, and we begin by analyzing its absolute moment bound. Our analysis techniques are adapted from Freksen et al. (2018) (Lemma 3). In our setting, the random hash function takes values uniformly from the fourth roots of unity $\{1, \omega, \omega^2, \omega^3\}$, rather than the real Rademacher set $\{1, -1\}$ considered in Freksen et al. (2018). Using the moment bound, we obtain an $(\varepsilon, \delta)$ norm preservation guarantee for the Complex `CountSketch`. Specifically, for any $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$, the sketch dimension $D$ can be chosen sufficiently large such that $\Pr\big[\big|\|\mathbf{Cx}\|_2^2 - \|\mathbf{x}\|_2^2\big| \geq \varepsilon \|\mathbf{x}\|_2^2\big] \leq \delta$, which is shown in Theorem 12. Furthermore, in Corollaries 13 and 14, we provide proofs for approximate matrix multiplication and spectral approximation of the Gram matrix, respectively. We then extend these bound and guarantees to the Complex `TensorSketch` in Theorem 15 by leveraging the fact that `TensorSketch` matrix can be viewed as a `CountSketch` matrix (Section 4 of Pham & Pagh (2013)). We begin with the following notation which we used in our analysis.

**Notation 1.** *For every $D, t, k > 0$, define*

$$\Lambda(D,t,k) = \begin{cases} \sqrt{\dfrac{t}{D}}, & k \geq Dt, \\ \max\left\{ \sqrt{\dfrac{t}{D}},\ \dfrac{t^2}{k \ln^2\!\left(\dfrac{\epsilon Dt}{k}\right)} \right\}, & Dt > k \geq \sqrt{Dt}, \\ \max\left\{ \sqrt{\dfrac{t}{D}},\ \dfrac{t^2}{k \ln^2\!\left(\dfrac{\epsilon Dt}{k}\right)},\ \dfrac{t}{k \ln\!\left(\dfrac{\epsilon Dt}{k^2}\right)} \right\}, & k < \sqrt{Dt}. \end{cases} \tag{175}$$

**Theorem 11.** *[Absolute moment bound for `CountSketch`] Let $D \in \mathbb{N}$ and let $t \leq D/4$ be an even integer. For any vector $\mathbf{x} \in \mathbb{R}^d$, let $\mathbf{C} \in \mathbb{C}^{D \times d}$ denote a Complex `CountSketch` matrix. Then,*

$$\big\|\|\mathbf{Cx}\|_2^2 - \|\mathbf{x}\|_2^2\big\|_{L^t} \;=\; O\big(\Lambda(D,\, t,\, k)\ \|\mathbf{x}\|_2^2\big)\,, \tag{176}$$

*where, $k = \|\mathbf{x}\|_\infty^{-2}$ denotes the inverse squared $\ell_\infty$-norm of the input vector, and $\Lambda(\cdot)$ is defined in Notation 1, where $\|X\|_{L^t} := (\mathbb{E}[|X|^t])^{1/t}$*

*Proof.* The detailed proof is given in Appendix A.1. □

We now provide an $(\varepsilon, \delta)$ norm preservation guarantee for the complex `CountSketch` using the absolute moment bound established above.

**Theorem 12** (Norm preservation for Complex `CountSketch`)**.** *Let $0 < \varepsilon < 1$ and $0 < \delta < e^{-2}$. For any vector $\mathbf{x} \in \mathbb{R}^d$, let $\mathbf{C} \in \mathbb{C}^{D \times d}$ denote a* Complex `CountSketch` *matrix as defined in Definiton 4.1. Then, with probability at least $1 - \delta$, we have,*

$$\big|\|\mathbf{Cx}\|_2^2 - \|\mathbf{x}\|_2^2\big| \;\leq\; \varepsilon \|\mathbf{x}\|_2^2, \tag{177}$$

*provided that the sketching dimension satisfies*

$$D = \Omega\left(\max\left\{\varepsilon^{-2},\ \varepsilon^{-1}\|\mathbf{x}\|_\infty^2 \log\left(\tfrac{1}{\delta}\right)\right\} \log\left(\tfrac{1}{\delta}\right)\right). \tag{178}$$

*Proof.* We first outline the structure of the proof. We define the error variable $X = \|\mathbf{Cx}\|_2^2 - \|\mathbf{x}\|_2^2$, which has zero mean by unbiasedness of the Complex `CountSketch`. Then by applying Markov's inequality on absolute moment bound in Theorem 11, $\|X\|_{L^t}$ provides a tail bound with $t = \Theta(\log(1/\delta))$. Analyzing the three regimes of $\Lambda(D, t, k)$ gives the lower bound on $D$ which ensures $(\epsilon, \delta)$ norm preservation guarantee.

We now present the detailed proof. We start by introducing the error variable.

$$X \;:=\; \|\mathbf{Cx}\|_2^2 - \|\mathbf{x}\|_2^2, \tag{179}$$

which measures the error in the estimation of squared $\ell_2$-norm. For a Complex `CountSketch` matrix, Theorem 1 establishes that the estimator $\|\mathbf{Cx}\|_2^2$ is unbiased, i.e.,

$$\mathbb{E}\big[\|\mathbf{Cx}\|_2^2\big] \;=\; \|\mathbf{x}\|_2^2. \tag{180}$$

Consequently,

$$\mathbb{E}[X] \;=\; \mathbb{E}\big[\|\mathbf{Cx}\|_2^2\big] - \|\mathbf{x}\|_2^2 \;=\; 0. \tag{181}$$

Let $t \le D/4$ be an even integer. By the absolute moment bound for the `CountSketch` matrix (Theorem 11), we have

$$\|X\|_{L^t} \;=\; \big(\mathbb{E}[|X|^t]\big)^{1/t} \;=\; O\big(\Lambda(D,t,k)\,\|\mathbf{x}\|_2^2\big)\,, \tag{182}$$

where $k \;=\; \|\mathbf{x}\|_\infty^{-2}$. Now by applying Markov's inequality to the non-negative random variable $|X|^t$ gives

$$\Pr\big(|X| \ge \varepsilon\|\mathbf{x}\|_2^2\big) = \Pr\big(|X|^t \ge (\varepsilon\|\mathbf{x}\|_2^2)^t\big)\,, \tag{183}$$

$$\le \frac{\mathbb{E}[|X|^t]}{\varepsilon^t\|\mathbf{x}\|_2^{2t}}. \tag{184}$$

Putting bound of $\mathbb{E}[|X|^t]$ using Equation (182), we get

$$\Pr\big(|X| \ge \varepsilon\|\mathbf{x}\|_2^2\big) \le \frac{c\big(\,\Lambda(D,t,k)\,\|\mathbf{x}\|_2^2\big)^t}{\varepsilon^t\|\mathbf{x}\|_2^{2t}}, \tag{185}$$

$$= c\left(\frac{\Lambda(D,t,k)}{\varepsilon}\right)^t. \tag{186}$$

where $c > 0$ is an absolute constant. Our objective is to make the right-hand side at most $\delta$. We choose $t := \big\lceil \log\big(\frac{1}{\delta}\big)\big\rceil$ , which ensures that $e^{-t} \le \delta$. It therefore guarantee that

$$\left(\frac{\Lambda(D,t,\|\mathbf{x}\|_\infty^{-2})}{\varepsilon}\right)^t \;\le\; e^{-t}, \tag{187}$$

whenever the base of the exponent is bounded by a constant $c \ge e$, then

$$\Lambda(D,t,\|\mathbf{x}\|_\infty^{-2}) \;\le\; \tfrac{\varepsilon}{c}, \tag{188}$$

which gives,

$$\Pr\big(|X| \ge \varepsilon\|\mathbf{x}\|_2^2\big) \;\le\; \big(\tfrac{1}{c}\big)^t = e^{-t\log c} \;\le\; e^{-t} \;\le\; \delta. \tag{189}$$

Finally, the constraint $t \le D/4$ is required in order to apply Theorem 11. We verify (189) by considering the three regimes in the definition of $\Lambda(D,t,k)$ (Notation 1).

*Case I: ($k \ge Dt$).* In this case,

$$\Lambda(D,t,k) = \sqrt{\frac{t}{D}}. \tag{190}$$

Thus $\Lambda(D,t,k) \le \frac{\varepsilon}{c}$ whenever

$$D \;\ge\; \frac{c^2 t}{\varepsilon^2}. \tag{191}$$

*Case II: ($\sqrt{Dt} \le k < Dt$).* Recall that

$$\Lambda(D,t,k) = \max\left\{\sqrt{\frac{t}{D}},\ \frac{t^2}{k\ln^2\left(\frac{\varepsilon Dt}{k}\right)}\right\}. \tag{192}$$

The first term is bounded exactly as in Case I, which gives the condition $D \ge c\,\frac{t}{\varepsilon^2}$. Then, we bound the second term by ensuring that,

$$\frac{t^2}{k\ln^2\left(\frac{\varepsilon Dt}{k}\right)} \ \le\ \frac{\varepsilon}{c}. \tag{193}$$

Since $t = \Theta(\log(1/\delta))$ and $D \ge ct$, and noting that $k < Dt$ in this regime, we have

$$\ln\left(\frac{\varepsilon Dt}{k}\right) = \Omega(\ln(Dt)). \tag{194}$$

Therefore, choosing

$$D\ \ge\ c\,\frac{t^2}{k\varepsilon}, \tag{195}$$

ensures that (193) holds. Combining this condition given in Equation (195) with bound in Equation (191) which provides condition,

$$D\ \ge\ c\max\left\{\frac{t}{\varepsilon^2},\ \frac{t^2}{k\varepsilon}\right\}. \tag{196}$$

*Case III: ($k < \sqrt{Dt}$).* In this case,

$$\Lambda(D,t,k) = \max\left\{\sqrt{\frac{t}{D}},\ \frac{t^2}{k\ln^2\left(\frac{\epsilon Dt}{k}\right)},\ \frac{t}{k\ln\left(\frac{\epsilon Dt}{k^2}\right)}\right\}. \tag{197}$$

The first two terms are bounded similarly as in Case II. We now analyze for the third term by ensuring that

$$\frac{t}{k\ln\left(\frac{\varepsilon Dt}{k^2}\right)} \ \le\ \frac{\varepsilon}{c}. \tag{198}$$

Since $t = \Theta(\log(1/\delta))$ and $D \ge ct$, and noting that $k^2 < Dt$ in this regime, we have

$$\ln\left(\frac{\varepsilon Dt}{k^2}\right) = \Omega(\ln(Dt)). \tag{199}$$

Therefore, choosing

$$D\ \ge\ c\,\frac{t}{k\varepsilon}. \tag{200}$$

ensures that the logarithmic denominator in (198) is sufficiently large, and hence the inequality holds. Combining the three regimes, we conclude that

$$\begin{aligned} D\ &\ge\ c\max\left\{\frac{t}{\varepsilon^2},\ \frac{t}{\varepsilon k},\ \frac{t^2}{\varepsilon k}\right\}, && (201)\\ &= c\max\left\{\frac{t}{\varepsilon^2},\ \frac{t^2}{\varepsilon k}\right\}, && (202)\\ &= c\max\left\{\varepsilon^{-2},\ \varepsilon^{-1}\|\mathbf{x}\|_\infty^2\log\left(\tfrac{1}{\delta}\right)\right\}\log\left(\tfrac{1}{\delta}\right), && (203) \end{aligned}$$

is sufficient to guarantee Equation (189). Therefore,

$$\Pr\big[\big|\|\mathbf{Cx}\|_2^2 - \|\mathbf{x}\|_2^2\big| \le \varepsilon\|\mathbf{x}\|_2^2\big] \;\ge\; 1-\delta, \tag{204}$$

provided that

$$D = \Omega\left(\max\big\{\varepsilon^{-2},\; \varepsilon^{-1}\|\mathbf{x}\|_\infty^2 \log\big(\tfrac{1}{\delta}\big)\big\}\log\big(\tfrac{1}{\delta}\big)\right). \tag{205}$$

□

Using the norm preservation, we can obtaing following approximate matrix product guarantee.

**Corollary 13** (Approximate Matrix Product)**.** *Let $0 < \varepsilon < 1$ and $0 < \delta < \exp(-2)$. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{d_1\cdots d_p}$ and let $\mathbf{C}$ be defined as in Theorem 4.1. In order to guarantee*

$$Pr\big[\big|\langle \mathbf{Cx}, \overline{\mathbf{Cy}}\rangle - \mathbf{x}^\top\mathbf{y}\big| \;\le\; \varepsilon\,\|\mathbf{x}\|\,\|\mathbf{y}\|\big] \;\ge\; 1-\delta, \tag{206}$$

*or, for two matrices $\mathbf{X} \in \mathbb{R}^{d_1\cdots d_p\times n}$ and $\mathbf{Y} \in \mathbb{R}^{d_1\cdots d_p\times m}$,*

$$Pr\left[\frac{\Big\|(\mathbf{CX})^\top\overline{(\mathbf{CY})} - \mathbf{X}^\top\mathbf{Y}\Big\|_F}{\|\mathbf{X}\|_F\,\|\mathbf{Y}\|_F} \;\le\; \varepsilon\right] \;\ge\; 1-\delta, \tag{207}$$

*the sketching matrix $\mathbf{C}$ must have $D$ rows, where $D$ is the same as in Theorem 12.*

*Proof.* The proof is deferred to Appendix A.2. □

We next apply the approximate matrix product guarantee to obtain a spectral approximation for the Gram matrix.

**Corollary 14** (Spectral Approximation of the Gram Matrix)**.** *Let $\mathbf{K} := (\mathbf{A}^{\otimes p})^\top\mathbf{A}^{\otimes p} \in \mathbb{R}^{n\times n}$ denote the Gram matrix corresponding to the points $\{\mathbf{a}_i\}_{i=1}^n$, and let*

$$\widehat{\mathbf{K}} \;:=\; (\mathbf{CA}^{\otimes p})^\top\overline{(\mathbf{CA}^{\otimes p})} \tag{208}$$

*be its randomized approximation. Then, with probability at least $1-\delta$, we have*

$$(1-\varepsilon)\,(\mathbf{K}+\lambda\mathbf{I}) \;\preceq\; \widehat{\mathbf{K}}+\lambda\mathbf{I} \;\preceq\; (1+\varepsilon)\,(\mathbf{K}+\lambda\mathbf{I}), \tag{209}$$

*for some $\lambda \ge 0$, provided that $\mathbf{C}$ has $2D\,s_\lambda(\mathbf{K})^2$ rows, where $D$ is the same as in Theorem 12 and*

$$s_\lambda(\mathbf{K}) \;:=\; \mathrm{Tr}\big(\mathbf{K}(\mathbf{K}+\lambda\mathbf{I})^{-1}\big) \;\le\; n \tag{210}$$

*is the $\lambda$-statistical dimension of $\mathbf{K}$.*

*Proof.* The proof is deferred to Appendix A.3. □

We now extend the above guaranties from Complex `CountSketch` to Complex `TensorSketch`.

**Theorem 15** (Guarantees for `TensorSketch`)**.** *Let $p \ge 1$ be an integer and let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{d_1\cdots d_p}$. Let $\mathbf{C} \in \mathbb{C}^{D\times d_1\cdots d_p}$ denote the Complex `TensorSketch` matrix as stated in Definition 4.2. Then we have the following:*

***(i) Absolute moment bound.*** *Let $t \le D/4$ be an even integer and define $k := \|\mathbf{x}\|_\infty^{-2}$. Then,*

$$\big\|\|\mathbf{Cx}\|_2^2 - \|\mathbf{x}\|_2^2\big\|_{L^t} = O\big(\Lambda(D,t,k)\,\|\mathbf{x}\|_2^2\big). \tag{211}$$

***(ii) Norm preservation.*** *For any $0 < \varepsilon < 1$ and $0 < \delta < e^{-2}$, with probability at least $1-\delta$,*

$$\big|\|\mathbf{Cx}\|_2^2 - \|\mathbf{x}\|_2^2\big| \le \varepsilon\,\|\mathbf{x}\|_2^2, \tag{212}$$

*provided that*

$$D = \Omega\left(\max\left\{\varepsilon^{-2}, \varepsilon^{-1}\left\|\mathbf{x}\right\|_\infty^2\right\}\log\left(\tfrac{1}{\delta}\right)\right). \tag{213}$$

***(iii) Approximate matrix product.*** *For any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{d_1 \cdots d_p}$,

$$Pr\big[\big|\langle \mathbf{Cx}, \overline{\mathbf{Cy}}\rangle - \langle \mathbf{x}, \mathbf{y}\rangle\big| \le \varepsilon \|\mathbf{x}\|_2 \|\mathbf{y}\|_2\big] \ge 1 - \delta. \tag{214}$$

*Moreover, for matrices* $\mathbf{X} \in \mathbb{R}^{d_1 \cdots d_p \times n}$ *and* $\mathbf{Y} \in \mathbb{R}^{d_1 \cdots d_p \times m}$,

$$Pr\left[\frac{\left\|(\mathbf{CX})^\top \overline{(\mathbf{CY})} - \mathbf{X}^\top \mathbf{Y}\right\|_F}{\|\mathbf{X}\|_F \|\mathbf{Y}\|_F} \le \varepsilon\right] \ge 1 - \delta. \tag{215}$$

***(iv) Spectral approximation of the Gram matrix.*** *Let* $\mathbf{K} := (\mathbf{A}^{\otimes p})^\top \mathbf{A}^{\otimes p} \in \mathbb{R}^{n \times n}$ *and*

$$\widehat{\mathbf{K}} := (\mathbf{CA}^{\otimes p})^\top \overline{(\mathbf{CA}^{\otimes p})}. \tag{216}$$

*Then, with probability at least* $1 - \delta$,

$$(1 - \varepsilon)(\mathbf{K} + \lambda \mathbf{I}) \preceq \widehat{\mathbf{K}} + \lambda \mathbf{I} \preceq (1 + \varepsilon)(\mathbf{K} + \lambda \mathbf{I}), \tag{217}$$

*provided that* $\mathbf{C}$ *has* $2D\, s_\lambda(\mathbf{K})^2$ *rows, where*

$$s_\lambda(\mathbf{K}) = \mathrm{Tr}\big(\mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1}\big) \;\le\; n. \tag{218}$$

*Proof.* We establish the same guarantees for Complex `TensorSketch` as for Complex `CountSketch` by viewing the former as a `CountSketch` with aggregated functions $H(\mathbf{i}) = \sum_{j=1}^p h_j(i_j) \bmod D$ and $S(\mathbf{i}) = \prod_{j=1}^p s_j(i_j)$, where $\mathbf{i} = (i_1, i_2, \ldots, i_p)$ is a multi-index. Then $H(\cdot)$ and $S(\cdot)$ together define a Complex `CountSketch` applied to $\mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p} \in \mathbb{R}^{d^p}$, as also observed for their real-counterpart in Section 4 of Pham & Pagh (2013). Hence, Complex `TensorSketch` $\mathbf{C}$ is distributionally identical to Complex `CountSketch`, and all moment, norm preservation, matrix product, and spectral approximation guarantees follow directly from the corresponding Complex `CountSketch` results.

□

## 5 Experiments

The experiments section is organized as follows. Subsection 5.1 describes the datasets and evaluation metrics. Subsections 5.2, 5.3, and 5.4 present the performance of our proposed methods Complex `CountSketch`, Complex `TensorSketch`, and Complex `Recursive TensorSketch`, respectively and compare them against their corresponding baseline algorithms.

### 5.1 Experimental Setup

- **Datasets.** We evaluate the proposed sketching methods on both synthetic and real–world datasets in order to assess their behavior under controlled conditions as well as on practical data distributions.

  *Synthetic data.* For the synthetic experiments, we generate $n = 300$ random vectors in $\mathbb{R}^d$ with $d = 5$. Each coordinate is drawn independently from a standard Gaussian distribution, and all vectors are $\ell_2$–normalized. We consider polynomial kernels of degrees $p \in \{5, 10, 15\}$ and vary the sketch dimension as $D \in \{100, 200, 400\}$. This controlled setting isolates the effect of sketching randomness and polynomial degree without confounding structure from real data.

  *Real–world data.* We further evaluate all methods on several real world datasets with diverse characteristics: (i) MAGIC Gamma Telescope Bock (2004), with $d = 10$ real–valued features, (ii) ISOLET Cole & Fanty (1991), a speech recognition dataset with $d = 617$ features, (iii) COD–RNA Uzilov

et al. (2006), a biological dataset with $d = 8$ numerical attributes, and (iv) MNIST LeCun et al. (1998), a handwritten digit image dataset with $d = 784$ numerical attributes. All inputs are treated as real–valued feature vectors and $\ell_2$–normalized prior to sketching. For each dataset, we subsample up to $n = 3000$ data points (or fewer when the dataset size is smaller). Each configuration (fixed dataset, degree $p$, and sketch dimension $D$) is repeated over 20–100 independent random trials depending on the experiment (setting $p = 1$ typically uses 100 trials, while higher–degree tensor and recursive sketch experiments use 20 trials). For kernel approximation and timing experiments, we report averages over these repetitions to ensure statistical stability of both error and runtime measurements.

- **Baselines.** We compare the performance of the proposed complex sketching methods against the following baseline algorithms across all experiments. These baselines cover both dense Johnson–Lindenstrauss (JL)–type polynomial sketches and input–sparsity tensor sketching schemes, enabling a systematic evaluation of variance and runtime complexities:
  - **Real Gaussian/Rademacher Sketch Kar & Karnick (2012):** A dense real Johnson–Lindenstrauss–type polynomial sketch using i.i.d. Gaussian projection matrices or using i.i.d. Rademacher projection matrices.
  - **Real `CountSketch` Charikar et al. (2004):** The classical `CountSketch` with real random signs, serving as the primary input–sparsity baseline for inner–product and polynomial kernel estimation.
  - **Real `TensorSketch` Pham & Pagh (2013):** The standard CountSketch–based `TensorSketch` construction for polynomial kernels, using real hash and sign functions.
  - **Real `Recursive TensorSketch` Ahle et al. (2020) :** This framework based on Real `CountSketch` and `TensorSketch` of degree-2 blocks.
  - **Complex Gaussian/Rademacher Sketch Wacker et al. (2024):** A dense complex Johnson–Lindenstrauss–type polynomial sketch using complex Gaussian projections or sing complex Rademacher projections constructed from from paired real Rademacher vectors.
  - **Complex `CountSketch` (Section 4.1) / `TensorSketch` (Section 4.2) / `Recursive TensorSketch` (Section 4.3) :** Our complex variants obtained by replacing real random signs with random hash function whose values are drawn independently and uniformly from the four fourth roots of unity within `CountSketch`, `TensorSketch`, and `Recursive TensorSketch` constructions.
- **Comparison Metrics.** We evaluate approximation quality using the following metrics.
  - *KL divergence.* To measure how well a sketch preserves the global structure of the polynomial kernel matrix, we compute the Kullback–Leibler (KL) divergence between the exact kernel matrix $K$ and its approximation $\widehat{K}$ obtained from the sketch. Since kernel matrices are nonnegative for polynomial kernels, we first normalize them into discrete distributions:

$$P_{ij} = \frac{K_{ij}}{\sum_{a,b} K_{ab}}, \qquad Q_{ij} = \frac{\widehat{K}_{ij}}{\sum_{a,b} \widehat{K}_{ab}}.$$

    The KL divergence is then defined as

$$\mathrm{KL}(K \,\|\, \widehat{K}) = \sum_{i,j} P_{ij} \log\left(\frac{P_{ij}}{Q_{ij} + \varepsilon}\right),$$

    where $\varepsilon > 0$ is a small constant for numerical stability.
  - *Wall–clock time.* We measure computational efficiency using wall–clock sketch construction time. For each method and sketch dimension $D$, we record the elapsed runtime required to construct the sketch features (or sketch vectors) and report the average over multiple independent trials. This metric directly reflects the practical runtime behavior and highlights the difference between dense JL–type projections and input–sparsity sketching methods such as `CountSketch`, `TensorSketch`, and `Recursive TensorSketch`.

- **Machine Configuration.** All experiments are performed on a machine running Ubuntu 22.04.4 with an Intel® Core™ i9–14900K processor (24 cores, 32 threads) and 32 GB RAM. Unless otherwise stated, sketch dimensions are varied over a logarithmic grid, and wall–clock times are averaged over multiple trials. All sketching methods are implemented within the same framework to ensure fair comparisons.

### 5.2 Complex `CountSketch` Insights

We first present results for our proposed Complex `CountSketch` in the case $p = 1$ and compare it with the baseline algorithms. Our experiments demonstrate a clear advantage of our approach in terms of reduced variance, while retaining the input-sparsity dependent running time of the baseline methods.

We summarize the variance comparison between our proposed method and the baseline algorithms in Figures 1 and 3 for the MAGIC and ISOLET datasets, respectively. For both datasets, our proposed Complex `CountSketch` achieves the smallest variance among all baselines and closely matches the variance performance of dense complex JL–type estimators, particularly complex Rademacher projections . The remaining two baselines exhibit substantially larger variance. These empirical observations are consistent with our theoretical findings summarized in Table 4.1.

We summarize the empirical runtime comparison among the baseline methods in Figures 2 and 4 for the MAGIC and ISOLET datasets, respectively. Our results show that the proposed method achieves the fastest runtime, matching that of `CountSketch`. This behavior is consistent with its input-sparsity–dependent time complexity, as sketch construction scales linearly with the number of nonzero entries rather than the ambient dimension. In contrast, dense complex JL sketches incur substantially higher runtime due to their fully dense projections. Overall, these results demonstrate that complex randomization provides significant variance reduction while preserving sparsity-dependent runtime advantages.

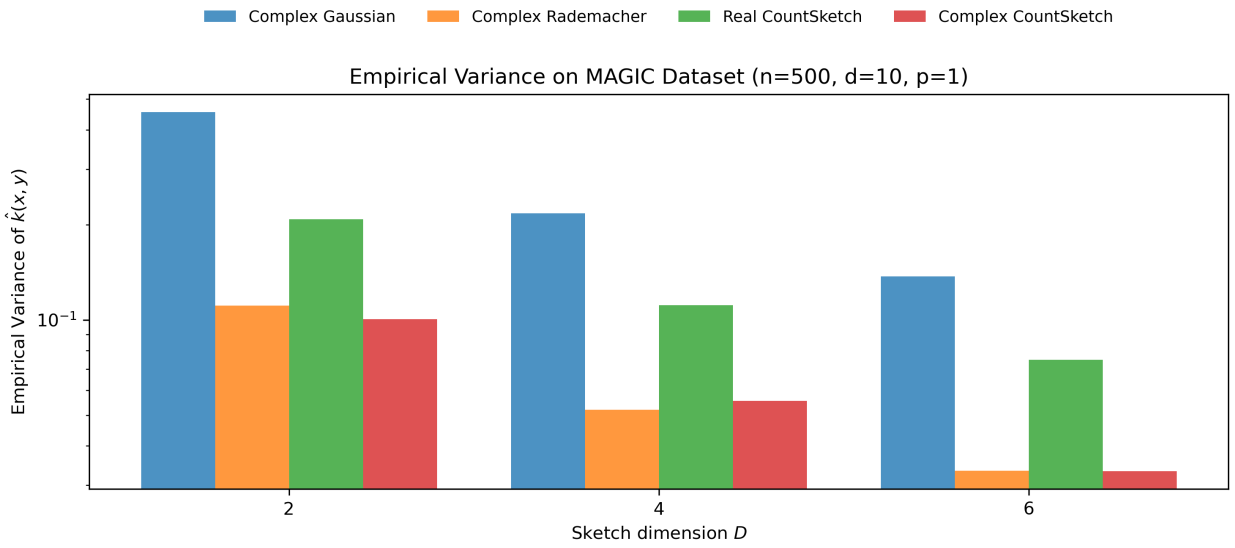


Figure 1: **Empirical variance comparison for inner product estimation ($p = 1$) on the MAGIC Telescope dataset.** We compare real and complex sketching methods Real CountSketch, Complex CountSketch, Complex Gaussian, and Complex Rademacher as a function of the sketch dimension $D$. Each bar reports the empirical variance of the estimator $\hat{k}(x, y)$ over 500 random trials, computed on randomly sampled normalized data points.

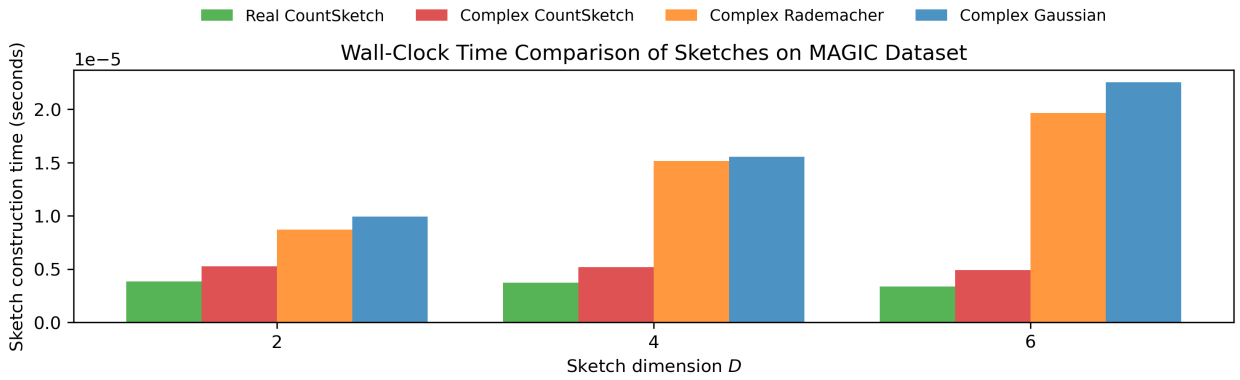


Figure 2: **Wall–clock sketch construction time on the MAGIC dataset ($p = 1$).** We report the average time required to construct a single sketch for Real CountSketch, Complex CountSketch, Complex Rademacher, and Complex Gaussian projections as a function of the sketch dimension $D$. All methods are evaluated on $\ell_2$–normalized MAGIC data with $n = 500$ samples, and timings are averaged over 100 independent runs.

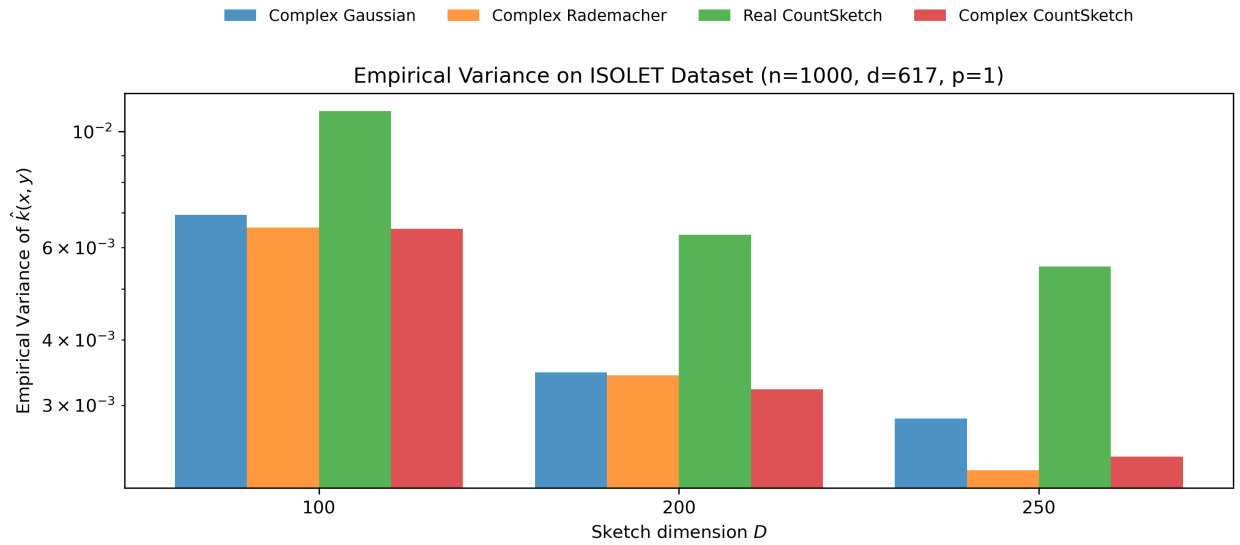


Figure 3: **Empirical variance comparison for inner product estimation ($p = 1$) on the ISOLET dataset.** We compare Real CountSketch, Complex CountSketch, Complex Gaussian, and Complex Rademacher sketches across increasing sketch dimensions $D$. The empirical variance of the estimator $\hat{k}(x, y)$ is computed over 100 independent trials using randomly sampled and $\ell_2$–normalized ISOLET dataset data points.

### 5.3 Complex `TensorSketch` Insights

We next present results for polynomial kernels of degree $p > 1$ and compare our proposed Complex `TensorSketch` with the baseline algorithms, including real and complex JL–type sketches (Gaussian and Rademacher) Kar & Karnick (2012); Wacker et al. (2024) and Real `TensorSketch` Pham & Pagh (2013). These experiments are conducted on both synthetic and real–world datasets and are given in Figures 5–10. The results show a consistent advantage of Complex `TensorSketch` over other baselines.

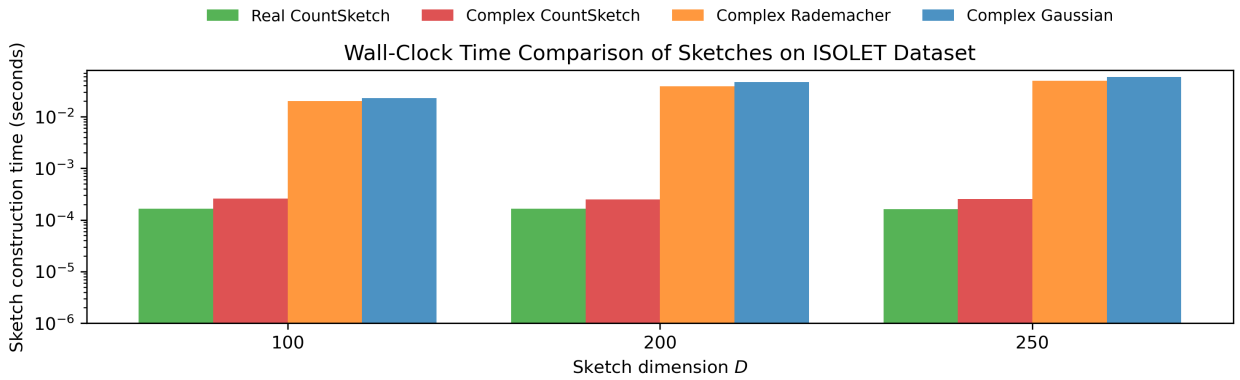


Figure 4: **Wall–clock sketch construction time on the ISOLET dataset ($p = 1$).** We compare the average time required to construct a single sketch for Real CountSketch, Complex CoWalluntSketch, Complex Rademacher, and Complex Gaussian projections as a function of the sketch dimension $D$. Reported times are averaged over 100 independent runs on randomly sampled and $\ell_2$–normalized data points.

We have given the variance comparison in Figures 5, 7, and 9. Across all datasets and polynomial degrees, our proposed Complex `TensorSketch` attains the lowest KL divergence among sparse sketching methods and closely matches the performance of dense complex JL sketches, particularly complex Rademacher projections. These empirical findings are consistent with the theoretical guarantees given in Table 4.2.

We have given the runtime comparison in Figures 6, 8, and 10. Our proposed Complex `TensorSketch` matches the time complexity of Real `TensorSketch` and remains significantly faster than all dense JL–type baselines across sketch dimensions and degrees. This behavior aligns with the input–sparsity time complexity of CountSketch-based constructions, where sketching cost depends primarily on the number of nonzeros rather than the ambient dimension. In contrast, real and complex Gaussian/Rademacher sketches incur substantially higher wall–clock time due to dense matrix multiplications. Just as we have seen in the previous section of the experiment, this complex-valued randomization achieves substantially lower variance while still maintaining the input-sparsity based computational efficiency.

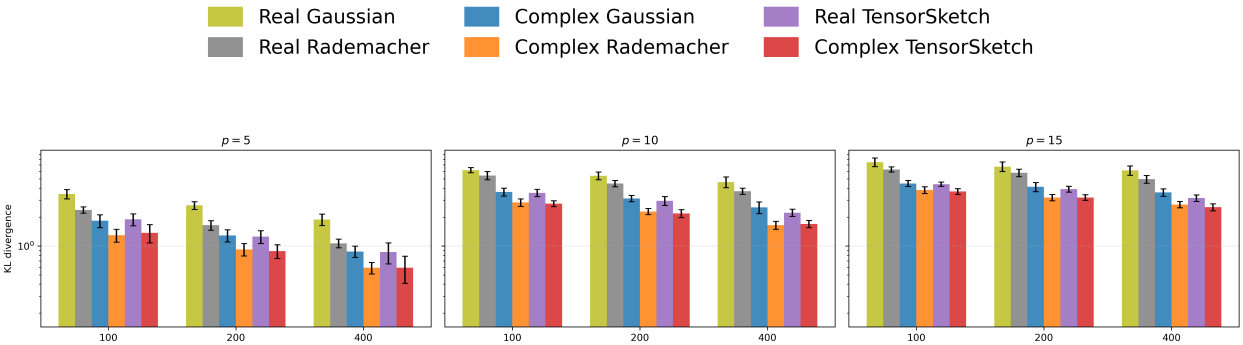


Figure 5: **KL divergence on synthetic dataset.** We report the KL divergence between the exact polynomial kernel matrix and its approximation obtained using different sketching methods, as a function of the sketch dimension $D$, for polynomial degrees $p \in \{5, 10, 15\}$. The synthetic dataset consists of $n = 300$ $\ell_2$–normalized vectors in dimension $d = 5$, and results are averaged over 20 independent trials.

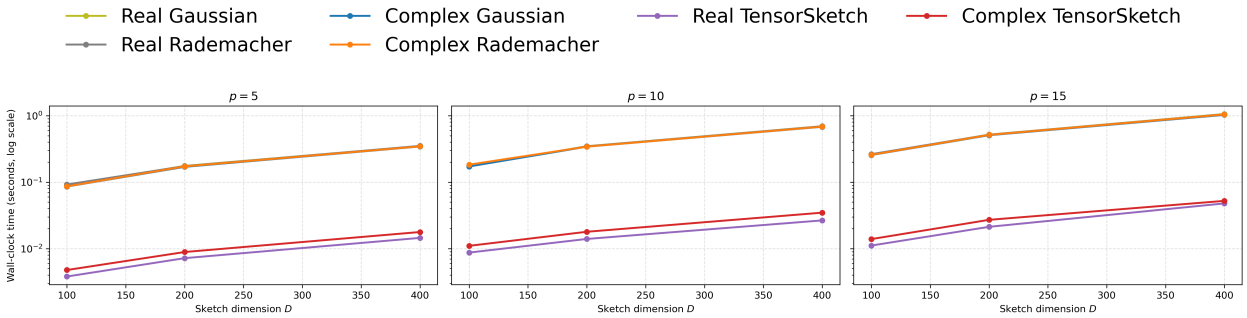


Figure 6: **Wall–clock time comparison on the synthetic dataset.** We report the average sketch construction time as a function of the sketch dimension $D$ for polynomial degrees $p \in \{5, 10, 15\}$. The synthetic dataset consists of $n = 300$ $\ell_2$–normalized vectors in dimension $d = 5$, and timings are averaged over multiple independent trials. Dense JL–type sketches (real and complex Gaussian/Rademacher) are compared against Real and Complex `TensorSketch`.

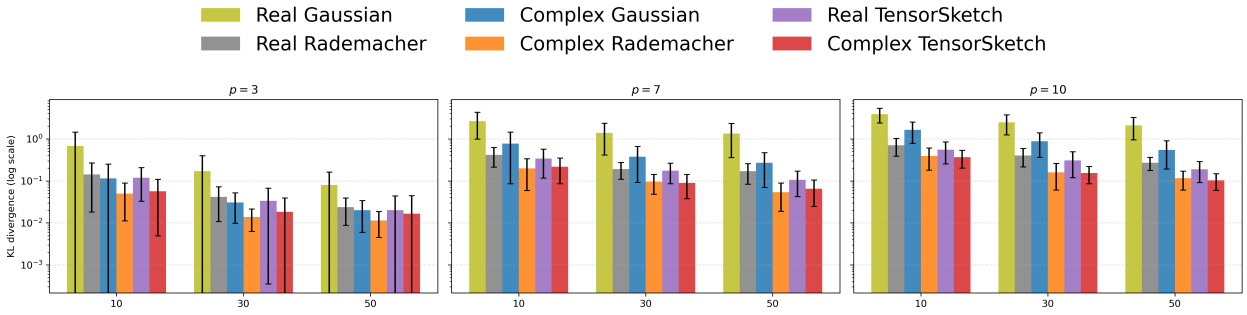


Figure 7: **KL divergence on the MAGIC Gamma Telescope dataset.** We compare the approximation quality of real and complex polynomial kernel sketches using the Kullback–Leibler (KL) divergence between the exact kernel matrix and its approximation. Results are shown for polynomial degrees $p \in \{3, 7, 10\}$ as a function of the sketch dimension $D \in \{d, 3d, 5d\}$, where $d = 10$ is the input dimension of the MAGIC dataset.

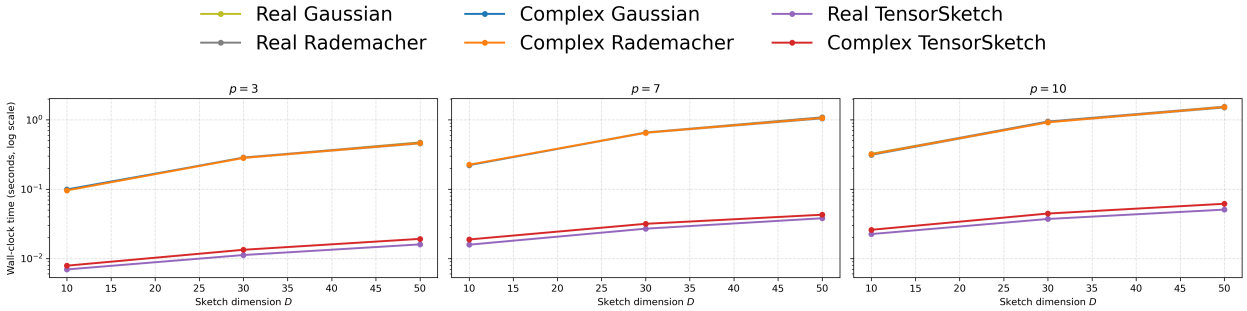


Figure 8: **Wall–clock time comparison on the MAGIC Gamma Telescope dataset.** We report the average sketch construction time (log scale) for real and complex polynomial sketching methods as a function of the sketch dimension $D \in \{d, 3d, 5d\}$, where $d = 10$ is the input dimension. Results are shown for polynomial degrees $p \in \{3, 7, 10\}$. Here as well all dense JL–type sketches exhibit nearly overlapping construction times across sketch dimensions $D$ and polynomial degrees $p$.

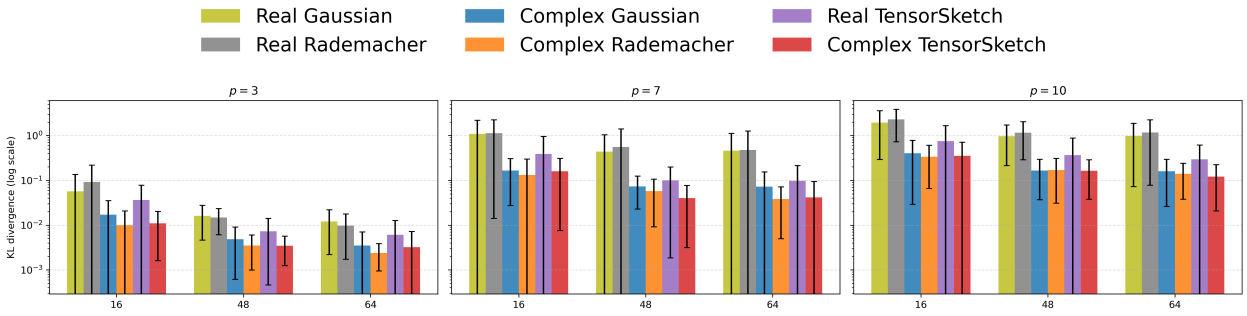


Figure 9: **KL divergence of polynomial kernel approximations on the COD–RNA dataset.** We compare real and complex JL–type sketches (Gaussian and Rademacher) with Real and Complex `TensorSketch` for polynomial degrees $p \in \{3, 7, 10\}$. The sketch dimension is varied as $D \in \{16, 48, 64\}$, corresponding to $D \in \{d, 3d, 5d\}$ with input dimension $d = 8$. Each bar reports the mean KL divergence between the exact kernel matrix $K$ and its approximation $\widehat{K}$ over 50 independent trials, with error bars indicating one standard deviation.

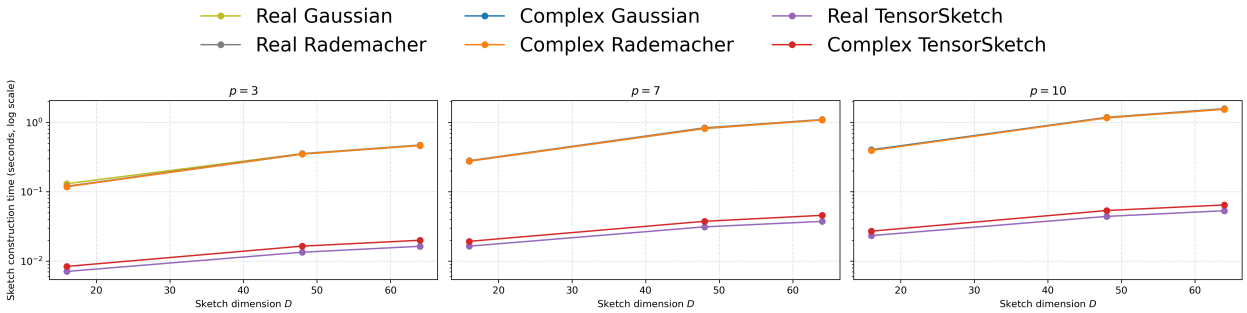


Figure 10: **Wall-clock sketch construction time on the COD–RNA dataset.** We compare real and complex JL–type polynomial sketches (Gaussian and Rademacher) with Real and Complex `TensorSketch` for polynomial degrees $p \in \{3, 7, 10\}$. The sketch dimension is varied as $D \in \{16, 48, 64\}$. Each point reports the average sketch construction time over 20 independent trials, measured on identical normalized input data. Here as well all dense JL-type sketches exhibit nearly overlapping construction times across sketch dimensions $D$ and polynomial degrees $p$.

### 5.4 Complex `Recursive TensorSketch` Insights

We finally turn to recursive sketching constructions and evaluate our proposed Complex `Recursive TensorSketch` against the Real `Recursive TensorSketch` Ahle et al. (2020) baseline. Experiments are conducted on both synthetic data and the real world dataset (MNIST). A consistent advantage of the complex variant is observed across sketch dimensions and polynomial degrees for variance. Also, the complex variant aligns with its real counterpart w.r.to. time complexity. Summary results are reported between Figures 11 - 14.

We conclude that Complex `Recursive TensorSketch` consistently achieves lower KL divergence than its real counterpart across all kernel degrees and sketch dimensions (Figures 11 and 13). The gap becomes more pronounced as the polynomial degree increases, indicating that complex randomization continues to suppress cross–term variance even under recursive tensor compositions. This aligns with the theoretical gaurantees summarized in Section 4.3.

In terms of time complexity, Complex and Real `Recursive TensorSketch` display closely matched wall–clock performance across datasets and parameter settings (Figures 12 and 14). Overall, the complex recursive construction delivers improved variance while matching the practical efficiency of the real recursive baseline.

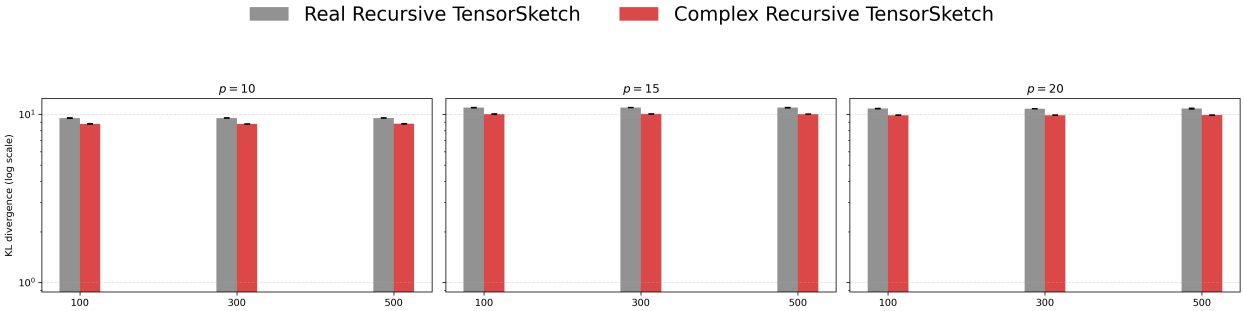


Figure 11: **KL divergence for Real vs. Complex `Recursive TensorSketch` on synthetic data.** We compare Real Recursive TensorSketch and Complex Recursive TensorSketch for polynomial kernel approximation on a synthetic dataset of $n = 1000$ $\ell_2$–normalized Gaussian vectors in dimension $d = 2$. Results are shown for polynomial degrees $p \in \{10, 15, 20\}$ and sketch dimensions $D \in \{100, 300, 500\}$. Each bar reports the mean KL divergence between the exact kernel matrix and its sketch-based approximation over 20 independent trials.

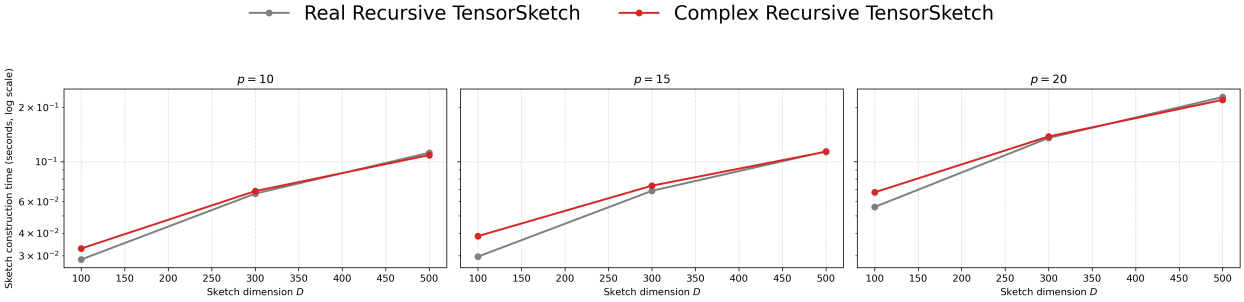


Figure 12: **Wall–clock time for Real vs. Complex `Recursive TensorSketch` on synthetic data.** We report the average sketch construction time for Real `Recursive TensorSketch` and Complex `Recursive TensorSketch` as a function of the sketch dimension $D$ for polynomial degrees $p \in \{10, 15, 20\}$. Experiments are conducted on a synthetic dataset of $n = 1000$ $\ell_2$–normalized Gaussian vectors in dimension $d = 2$, and each timing value is averaged over 20 independent runs.

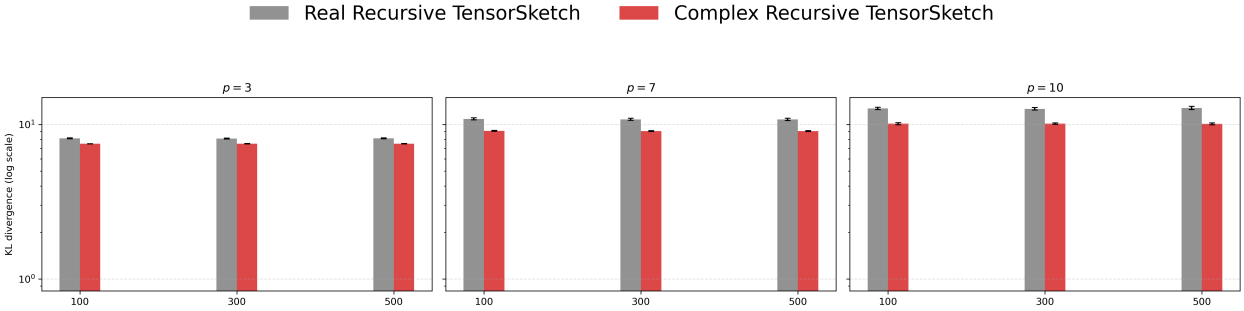


Figure 13: **KL divergence of Real vs. Complex `Recursive TensorSketch` on MNIST.** We compare Real `Recursive TensorSketch` and Complex `Recursive TensorSketch` for polynomial kernel approximation on the MNIST dataset. Results are shown for kernel degrees $p \in \{3, 7, 10\}$ and sketch dimensions $D \in \{100, 300, 500\}$. The dataset is $\ell_2$–normalized and restricted to $n = 1000$ samples, and each bar reports the mean KL divergence between the exact kernel matrix and its approximation over 20 independent trials.

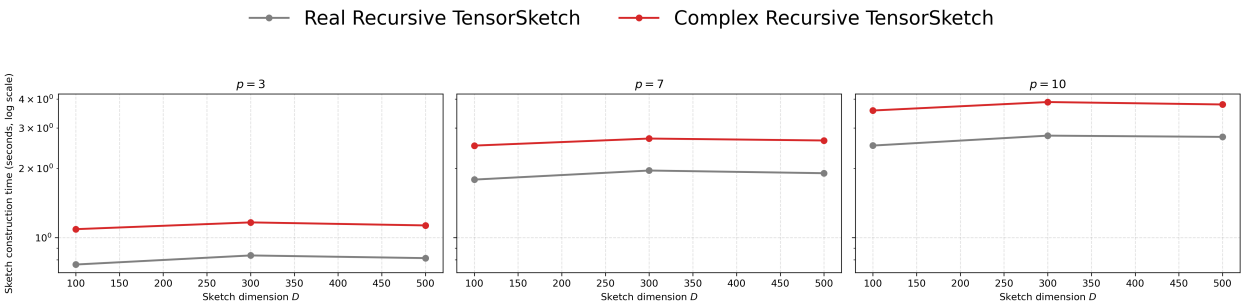


Figure 14: **Wall–clock sketch construction time for `Recursive TensorSketch` on MNIST.** We report the average time required to construct a single sketch using Real `Recursive TensorSketch` and Complex `Recursive TensorSketch` as a function of the sketch dimension $D \in \{100, 300, 500\}$ for polynomial degrees $p \in \{3, 7, 10\}$. Experiments are conducted on $\ell_2$–normalized MNIST data with $n = 1000$ samples, and timings are averaged over 20 independent runs.

## 6 Conclusion

In this paper, we introduced complex–valued variants of `CountSketch`, `TensorSketch`, and `Recursive TensorSketch` that retain the input–sparsity running time of their real–valued counterparts while achieving provably improved variance guarantees of their respective estimates. Our analysis shows that replacing real random signs with complex fourth–root–of–unity variables systematically removes cross–term contributions responsible for the higher variance observed in their respective real sketches. As a result, the proposed estimators achieve variance comparable to that of dense complex JL–type tensor estimators Wacker et al. (2024), while offering a running time that depends on the input sparsity, making our approach particularly well suited for high-dimensional sparse inputs. Beyond establishing these theoretical advantages, our constructions extend the scope of complex sketching to recursive sketching methods Ahle et al. (2020), demonstrating that the benefits of complex randomness are not limited to dense projection models.

Despite these advances, the following questions remain open:

- A limitation of our proposal is that the resultant sketches become complex vectors, which requires rendering a downstream task such as linear regression more involved due to linear algebra operations being applied to complex data. Recent work Wacker et al. (2023) studied real–valued $K_{\mathrm{CtR}}$ sketches constructed from complex–valued $K_C$ sketches Wacker et al. (2024), and showed that the variance advantages of complex estimators persist even after conversion to real sketches. Developing an analogous $K_{\mathrm{CtR}}$ style construction within our Complex `CountSketch` based framework remains an interesting open problem.
- While our results provide improved variance bounds with input–sparsity running time for Complex `Recursive TensorSketch`, establishing ($\varepsilon$–$\delta$) approximation guarantees (analogous to JL distortion bounds) remains an important open direction.
- Our analysis shows that using complex fourth roots of unity instead of real-valued random variables leads to improved variance. Understanding whether higher-order roots of unity can yield further improvements in higher-order moments remains an open question.

## References

Dimitris Achlioptas. Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003. doi: 10.1016/S0022-0000(03)00025-4.

Thomas D. Ahle, Michael Kapralov, Jakob Bæk Tejs Knudsen, Rasmus Pagh, Ameya Velingker, David P. Woodruff, and Amir Zandieh. Oblivious sketching of high-degree polynomial kernels. In *Proceedings of*

*the 31st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 141–160, Philadelphia, PA, 2020. SIAM. doi: 10.1137/1.9781611975994.9.

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson–Lindenstrauss transform. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 557–563, New York, NY, 2006. ACM.

Hélène Aschard. A perspective on interaction effects in genetic association studies. *Genetic Epidemiology*, 40(8):678–688, 2016. doi: 10.1002/gepi.21989.

R. Bock. MAGIC gamma telescope. UCI Machine Learning Repository, 2004. DOI: https://doi.org/10.24432/C52C8B.

Vladimir Braverman, Kai-Min Chung, Zhenming Liu, Michael Mitzenmacher, and Rafail Ostrovsky. AMS without 4-wise independence on product domains, 2010. URL `https://arxiv.org/abs/0806.4790`.

Moses Charikar, Kevin C. Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15, 2004. doi: 10.1016/S0304-3975(03)00400-6.

Wenlin Chen, James T. Wilson, Stephen Tyree, Kilian Q. Weinberger, and Yixin Chen. Compressing neural networks with the hashing trick. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 2285–2294, Lille, France, 2015. PMLR.

Krzysztof Marcin Choromanski, Mark Rowland, and Adrian Weller. The unreasonable effectiveness of structured random orthogonal embeddings. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 219–228, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/bf8229696f7a3bb4700cfddef19fa23f-Abstract.html`.

Kenneth L. Clarkson and David P. Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM*, 63(6):1–45, 2017. doi: 10.1145/3019134.

Ron Cole and Mark Fanty. ISOLET. UCI Machine Learning Repository, 1991. DOI: https://doi.org/10.24432/C51G69.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 3rd edition, 2009.

Casper B. Freksen, Lior Kamma, and Kasper Green Larsen. Fully understanding the hashing trick. *Advances in Neural Information Processing Systems*, 31, 2018.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 457–468, Austin, TX, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1044. URL `https://aclanthology.org/D16-1044/`.

Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 317–326, Los Alamitos, CA, 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.41.

Yoav Goldberg and Michael Elhadad. splitSVM: Fast, space-efficient, non-heuristic, polynomial kernel computation for NLP applications. In *Proceedings of ACL-08: HLT, Short Papers*, pp. 237–240, Columbus, OH, 2008. Association for Computational Linguistics. URL `https://aclanthology.org/P08-2060/`.

Ruhui Jin, Tamara G. Kolda, and Rachel A. Ward. Faster Johnson–Lindenstrauss transforms via Kronecker products. *CoRR*, abs/1909.04801, 2019. URL `http://arxiv.org/abs/1909.04801`.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984. URL `https://www.ams.org/books/conm/026/`.

Keegan Kang and Weipin Wong. Improving sign random projections with additional information. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pp. 2479–2487, Cambridge, MA, 2018. PMLR. URL https://proceedings.mlr.press/v80/kang18b.html.

Keegan Kang, Sergey Kushnarev, Wong Wei Pin, Rameshwar Pratap, Haikal Yeo, and Yijia Chen. Improving hashing algorithms for similarity search via MLE and the control variates trick. In *Proceedings of the 13th Asian Conference on Machine Learning (ACML)*, volume 157, pp. 814–829, Cambridge, MA, 2021. PMLR.

Purushottam Kar and Harish Karnick. Random feature maps for dot product kernels. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22, pp. 583–591, Cambridge, MA, 2012. JMLR.org. URL http://proceedings.mlr.press/v22/kar12.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Ping Li, Trevor Hastie, and Kenneth Ward Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 287–296, New York, NY, 2006. ACM. doi: 10.1145/1150402.1150436.

Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random Fourier features. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 3905–3914, Cambridge, MA, 2019. PMLR.

Raphael A. Meyer and Haim Avron. Hutchinson's estimator is bad at Kronecker-trace-estimation. *SIAM Journal on Matrix Analysis and Applications*, 47(1):353–387, 2026. doi: 10.1137/24M1720895.

Mihai Pătraşcu and Mikkel Thorup. The power of simple tabulation hashing. *Journal of the ACM*, 59(3): 1–50, 2012. doi: 10.1145/2213556.2213560.

Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 239–247, New York, NY, 2013. ACM. doi: 10.1145/2487575.2487591.

Ninh Pham and Rasmus Pagh. Tensor sketch: Fast and scalable polynomial kernel approximation. *CoRR*, abs/2505.08146, 2025. doi: 10.48550/ARXIV.2505.08146. URL https://doi.org/10.48550/arXiv.2505.08146.

Rameshwar Pratap and Raghav Kulkarni. Variance reduction in frequency estimators via control variates method. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 161, pp. 183–193, Corvallis, OR, 2021. AUAI Press. URL https://proceedings.mlr.press/v161/pratap21a.html.

Rameshwar Pratap, Bhisham Dev Verma, and Raghav Kulkarni. Improving Tug-of-War sketch using control-variates method. In *Proceedings of the 2021 SIAM Conference on Applied and Computational Discrete Algorithms (ACDA)*, pp. 66–76, Philadelphia, PA, 2021. SIAM. doi: 10.1137/1.9781611976830.7.

Steffen Rendle. Factorization machines. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM)*, pp. 995–1000, Los Alamitos, CA, 2010. IEEE Computer Society. doi: 10.1109/ICDM.2010.127.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2001.

Yitong Sun, Anna Gilbert, and Ambuj Tewari. But how does it work in theory? Linear SVM with random features. *Advances in Neural Information Processing Systems*, 31, 2018.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Andrew Uzilov, Joshua Keegan, and David Mathews. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, 7:173, 2006. doi: 10.1186/1471-2105-7-173.

Bhisham Dev Verma, Rameshwar Pratap, and Manoj Thakur. Variance reduction in feature hashing using MLE and control variate method. *Machine Learning*, 111(7):2631–2662, 2022. doi: 10.1007/S10994-022-06166-Z.

Bhisham Dev Verma, Punit Pankaj Dubey, Rameshwar Pratap, and Manoj Thakur. Improving compressed matrix multiplication using control variate method. *Information Processing Letters*, 187:106517, 2025. doi: 10.1016/J.IPL.2024.106517.

Jonas Wacker, Ruben Ohana, and Maurizio Filippone. Complex-to-real sketches for tensor products with applications to the polynomial kernel. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206, pp. 5181–5212, Cambridge, MA, 2023. PMLR. URL `https://proceedings.mlr.press/v206/wacker23a.html`.

Jonas Wacker, Motonobu Kanagawa, and Maurizio Filippone. Improved random features for dot product kernels. *Journal of Machine Learning Research*, 25:235:1–235:75, 2024. URL `https://jmlr.org/papers/v25/22-0118.html`.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 1113–1120, New York, NY, 2009. ACM.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Amir Zandieh, Insu Han, Haim Avron, Neta Shoham, Chaewon Kim, and Jinwoo Shin. Scaling neural tangent kernels via sketching and random features. *Advances in Neural Information Processing Systems*, 34:1062–1073, 2021.

# A Proofs of supporting lemmas

## A.1 Proof of Absolute moment bound for `CountSketch`

In this section, we prove Theorem 11. Our goal is to upper bound the moment of the Complex `CountSketch` error $\left|\|\mathbf{Cx}\|_2^2 - 1\right|$, where $\mathbf{C} \in \mathbb{C}^{D\times d}$ is a Complex `CountSketch` matrix that sketches unit vector $\mathbf{x} \in \mathbb{R}^d$ to dimension $D$. Our analysis adapts the proof technique of Freksen et al. (2018). While Freksen et al. (2018) derives bound for `CountSketch` with a real-valued sign hash function, in this work, we extend the analysis to the complex-valued function, which takes values independently and uniformly from fourth-root-of-unity $\{1, \omega, \omega^2, \omega^3\}$.

Formally, let $\alpha, \beta, t$ be integers such that $1 \le \beta \le \alpha/2 \le \min\{t/2,\, \alpha/2\}$. Let $\mathcal{G}_{\alpha,\beta,r}$ denote the family of all edge-labeled Eulerian multigraphs $G = ([\alpha], E_G, \pi_G)$, such that

1. $G$ has no isolated vertices;
2. $|E_G| = t$, and $\pi_G : E_G \to [t]$ is a bijection, which assigns a label in $[t]$ to each edge; and
3. the number of connected components in $G$ is $\beta$.

**Notation 2.** *Denote* $\Delta = \Delta(\alpha,\beta) = \alpha^{2\alpha-\beta}\Big[(\alpha-2\beta)^2 + 4(\alpha-\beta)\Big]^{r-\alpha}$.

**Corollary 16.** *(Freksen et al., 2018, Corollary 9)* $|\mathcal{G}_{\alpha,\beta,t}| \le 2^{O(t)}\Delta(\alpha,\beta)$.

Above corollary provides an upper bound on the number of labeled Eulerian graphs and can be used to bound the combinatorial terms appear in the moment analysis for the directed Eulerian graphs in our setting. Leveraging this estimate, we next bound the higher-order moments of the sketching error.

**Bounding the Moments of** $|\|\mathbf{Cx}\|_2^2 - 1|$. According to Definition 4.1, we know that $h \in_R [d] \to [D]$, and $s(1), \ldots, s(d) \in \{1, \omega, \omega^2, \omega^3\}$ are independent functions. The entries of the Complex `CountSketch` matrix are defined as $\mathbf{C}_{ij} := s(j)\cdot 1_{h(j)=i}, \forall i \in [D], j \in d$. Then we denote $X := \|\mathbf{Cx}\|_2^2 - 1$ for every unit vector $\mathbf{x} \in \mathbb{R}^d \setminus \{0\}$ . We start with providing a better understanding of $X$ as follows,

$$\|\mathbf{Cx}\|_2^2 = \langle \mathbf{Cx}, \overline{\mathbf{Cx}}\rangle = 1 + \sum_{j\neq\ell\in[n]} 1_{h(j)=h(\ell)}\cdot s(j)\overline{s(\ell)}\cdot x_j x_\ell, \tag{219}$$

Denote $[I]_p = \{(p,p) : p \in [d]\}$. Then

$$X = \left|\|\mathbf{Cx}\|_2^2 - 1\right| = \left|\sum_{(j,\ell)\in([d]\times[d]\setminus I_{[d]})} 1_{h(j)=h(\ell)}\cdot s(j)\overline{s(\ell)}\cdot x_j x_\ell\right|. \tag{220}$$

Hence, for every even $t$,

$$\|X\|_t^t = \mathbb{E}[X^t] = \sum_{\langle(j_p,\ell_p)\rangle_{p\in[t]}\in([d]\times[d]\setminus I_{[d]})^t} \mathbb{E}\left[\prod_{p\in[t]} 1_{h(j_p)=h(\ell_p)} s(j_p)\overline{s(\ell_p)} x_{j_p} x_{\ell_p}\right]. \tag{221}$$

Every $S = \langle(j_p,\ell_p)\rangle_{p\in[t]} \in ([d]\times[d]\setminus I_{[d]})^t$ defines a directed multigraph $G_S$ with $t$ ordered directed edges on vertex set $[d]$. The proof diverges from that of Freksen et al. (2018) starting here: while their analysis relies on undirected graph structures, our framework requires a directed graph formulation. We need the following definitions to adapt their proof outline in our setting:

**Definition 7** (In-degree and Out-degree)**.** Let $S = \langle(j_p,\ell_p)\rangle_{p\in[t]} \in ([d]\times[d]\setminus I_{[d]})^t$ be a sequence of ordered pairs over $[d]$, and let $G_S$ denote the directed graph induced by $S$ on the vertex set $[d]$. For any vertex $q \in [d]$, the *out-degree* and *in-degree* of $q$ in $S$ are defined as

- **Out-degree of $q$ :** $d_{S-\text{out}}(q) := |\{p \in [t] \ : \ j_p = q\}|$,
- **In-degree of $q$ :** $d_{S-\text{in}}(q) := |\{p \in [t] \ : \ \ell_p = q\}|$,
- **Degree of $q$ :** $d_S(q) := d_{S-\text{in}}(q) + d_{S-\text{out}}(q)$.

**Notation 3.** *Given $S \in ([d] \times [d] \setminus I_{[d]})^t$, let $CC(S)$ denote the set of all connected components of $G_S$ that contain at least two nodes. Let $\beta(S) := |CC(S)|$, $V(S) = \bigcup_{C \in CC(S)} C$, and $\alpha(S) := |V(S)|$.*

Now, for every integer $\beta$ and a subset $V \subseteq [d]$, let $\mathcal{S}_{V,\beta} \subseteq ([d] \times [d])^t$ be the set of all sequences $S \in ([d] \times [d])^t$ such that

1. For every $q \in [d]$, $d_S(q)$ is even,
2. $d_{S-\text{in}}(q) = d_{S-\text{out}}(q) = \frac{d_S(q)}{2}$,
3. $V(S) = V$ and $\beta(S) = \beta$.

For further analysis, we fix some $S = \langle (j_p, \ell_p) \rangle_{p \in [t]} \in ([d] \times [d] \setminus I_{[d]})^t$. Then continuing from Equation (221), we have the following

$$\begin{aligned}
&\mathbb{E}\left[\prod_{p\in[t]} 1_{h(j_p)=h(\ell_p)} s(j_p)\overline{s(\ell_p)} x_{j_p} x_{\ell_p}\right] \\
&\quad = \mathbb{E}\left[\prod_{p\in[t]} 1_{h(j_p)=h(\ell_p)}\right] \cdot \mathbb{E}\left[\prod_{q\in V(S)} s(q)^{d_{S-\text{in}}(q)} \overline{s(q)}^{d_{S-\text{out}}(q)}\right] \cdot \prod_{q\in V(S)} x_q^{d_S(q)}, \qquad (222) \\
&\quad = \mathbb{E}\left[\prod_{p\in[t]} 1_{h(j_p)=h(\ell_p)}\right] \cdot \mathbb{E}\left[\prod_{q\in V(S)} |s(q)|^{d_S(q)}\right] \cdot \prod_{q\in V(S)} x_q^{d_S(q)}, \qquad (223)
\end{aligned}$$

where the last equality follows from independence. Assume first that for some $q \in V(S)$, $d_S(q)$ is odd. Then $\mathbb{E}[s(q)^{d_S(q)}] = 0$, and therefore (223) equals 0. Otherwise, $\mathbb{E}[|s(q)|^{d_S(q)}] = 1$, $\mathbb{E}[|\sigma_q|^{d_S(q)}] = \mathbb{E}[s(q)^{\frac{d_S(q)}{2}} \overline{s(q)}^{\frac{d_S(q)}{2}}] = 1$, $\mathbb{E}[s(q)^{2a}] = 0$ and $\mathbb{E}[\overline{s(q)}^{2b}] = 0$ for all $q \in V(S)$ and $a, b \leq \frac{d_S(q)}{2}$. Hence, non-zero are where every vertex has an even degree with balanced in-degrees and out-degrees.

As mentioned earlier, in contrast, the analysis of Freksen et al. (2018) uses a hash function $s : [d] \to \{\pm 1\}$ satisfying $\mathbb{E}[s(q)^2] = 1$ . Our complex function $s : [d] \to \{1, \omega, \omega^2, \omega^3\}$ instead satisfies $\mathbb{E}[s(q)^2] = 0$ and $\mathbb{E}[\overline{s(q)}^2] = 0$ for all $q \in V(S)$, which eliminates all terms with unequal in-degree and out-degree.

For every $C \in CC(S)$, $C$ contains an edge of $G_S$, thus there exists $p \in [t]$ such that $j_p, \ell_p \in C$. Conversely, for every $p \in [t]$, there exists a unique connected component $C \in CC(S)$ such that $j_p, \ell_p \in C$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\prod_{p\in[t]} 1_{h(j_p)=h(\ell_p)}\right] &= \mathbb{E}\left[\prod_{C\in CC(S)} \prod_{p\in[t]:j_p\in C} 1_{h(j_p)=h(\ell_p)}\right], \qquad (224) \\
&= \prod_{C\in CC(S)} \mathbb{E}\left[\prod_{p\in[t]:j_p,\in C} 1_{h(j_p)=h(\ell_p)}\right], \qquad (225)
\end{aligned}$$

where the last equality is due to independence. Next, let $C = \{v_1, \ldots, v_{|C|}\} \in CC(S)$. We thus conclude that

$$\mathbb{E}\left[\prod_{p\in[t]} 1_{h(j_p)=h(\ell_p)}\right] = \prod_{C\in CC(S)} \frac{1}{D^{|C|-1}} = \frac{1}{D^{\alpha(S)-\beta(S)}}. \qquad (226)$$

For every sequence $S$ that donates a non-zero summand to the sum, since $d_S(q)$ is even for all $q \in V(S)$, every $C \in CC(S)$ is Eulerian, and therefore contains at least two nodes and two edges. Therefore $1 \leq \beta(S) \leq r/2$ and $2\beta(S) \leq \alpha(S) \leq r$. Plugging this into (221) we get that

$$\begin{aligned}
\|X\|_r^r &= \sum_{S \in ([d]\times[d]\setminus I_{[d]})^t} \mathbb{E}\left[\prod_{p\in[t]} 1_{h(j_p)=h(\ell_p)} s(j_p)\overline{s(\ell_p)} x_{j_p} x_{\ell_p}\right], \\
&= \sum_{\substack{S \in ([d]\times[d]\setminus I_{[d]})^t, \\ \forall q \in V(S).\, d_S(q) \in \mathbb{N}_{\text{even}}}} \frac{1}{D^{\alpha(S)-\beta(S)}} \prod_{q\in V(S)} x_q^{d_S(q)}, \qquad (227) \\
&= \sum_{\beta=1}^{t/2} \sum_{\alpha=2\beta}^{t} \sum_{V\in[d]} \sum_{S\in\mathcal{S}_{V,\beta}} \frac{1}{D^{\alpha-\beta}} \prod_{q\in V} x_q^{d_S(q)}. \qquad (228)
\end{aligned}$$

For every $q \in V(S)$, $d_S(q)$ is a positive even integer, and therefore $d_S(q)-2 \geq 0$ is also even. Hence for every $q \in V(S)$, $x_q^{d_S(q)} = |x_q|^{d_S(q)-2} \cdot |x_q|^2 \leq \|x\|_\infty^{d_S(q)-2} x_q^2$. Since $\sum_{q\in V(S)} d_S(q) = 2t$, then $\prod_{q\in V(S)} |x_q|^{d_S(q)} \leq \|x\|_\infty^{2t-2\alpha} \prod_{q\in V(S)} x_q^2$. Moreover, equality holds if for all $j \in \text{supp}(x)$, $|x_j| = \|x\|_\infty$. Plugging this in (228) we get that

$$\begin{aligned}
\|X\|_t^t &= \sum_{\beta=1}^{t/2} \sum_{\alpha=2\beta}^{t} \frac{D^\beta}{D^\alpha} \|\mathbf{x}\|_\infty^{2t-2\alpha} \sum_{V\in([d])^\alpha} |\mathcal{S}_{V,\beta}| \prod_{q\in V} x_q^2, \\
&= \|\mathbf{x}\|_\infty^{2t} \sum_{\beta=1}^{t/2} \sum_{\alpha=2\beta}^{t} \frac{D^\beta}{(D\|\mathbf{x}\|_\infty^2)^\alpha} \sum_{V\in([d])^\alpha} |\mathcal{S}_{V,\beta}| \prod_{q\in V} x_q^2. \qquad (229)
\end{aligned}$$

To upper bound the expression above equation, we use the following combinatorial claim and lemmas of Freksen et al. (2018), as it involves the same summation structure and combinatorial terms.

**Notation 4.** *for every $1 \leq \beta \leq \alpha/2 \leq t/2$ , define*

$$\begin{aligned}
M(\alpha,\beta) &= (D\beta^{-1})^\beta \left(\frac{k\alpha}{D}\right)^\alpha (\alpha-2\beta)^{2t-2\alpha}, \\
N(\alpha,\beta) &= (D\beta^{-1})^\beta \left(\frac{k\alpha}{D}\right)^\alpha (\alpha-\beta)^{t-\alpha}.
\end{aligned}$$

**Claim 17** (Freksen et al. (2018))**.**

$$\|X\|_t^t \;\leq\; \frac{2^{O(t)}}{k^t} \sum_{\beta=1}^{t/2} \sum_{\alpha=2\beta}^{t} \big(M(\alpha,\beta) + N(\alpha,\beta)\big). \qquad (230)$$

**Lemma 18.** *(Freksen et al., 2018, Lemma 18) For all $1 \leq \beta \leq \alpha/2 \leq t/2$, if $k \geq Dt$, then*

$$M(\alpha,\beta) \;\leq\; 2^{O(t)} \left(\frac{k^2 t}{D}\right)^{t/2}. \qquad (231)$$

*Otherwise,*

$$M(\alpha,\beta) \;\leq\; 2^{O(t)} \max\left\{ \left(\frac{t}{\ln \frac{2Dt}{k}}\right)^{2t}, \left(\frac{k^2 t}{D}\right)^{t/2} \right\}. \qquad (232)$$

**Lemma 19.** *(Freksen et al., 2018, Lemma 19) Let $1 \leq \beta \leq \alpha/2 \leq t/2$. If $k^2 > Dt$, then*

$$N(\alpha,\beta) \;\leq\; \left(\frac{k^2 t}{D}\right)^{t/2}. \qquad (233)$$

*Otherwise,*

$$N(\alpha,\beta) \;\le\; \max\left\{ M(\alpha,\beta), \left(\frac{t}{\ln\frac{Dt}{k^2}}\right)^t \right\}. \tag{234}$$

We now prove the main theorem, which upper bounds the moment of Complex `CountSketch` error using Claim 17, Lemma 18, and Lemma 19.

**Theorem 11.** *[Absolute moment bound for `CountSketch`] Let $D \in \mathbb{N}$ and let $t \le D/4$ be an even integer. For any vector $\mathbf{x} \in \mathbb{R}^d$, let $\mathbf{C} \in \mathbb{C}^{D\times d}$ denote a Complex `CountSketch` matrix. Then,*

$$\left\| \|\mathbf{C}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right\|_{L^t} \;=\; O\big(\Lambda(D,\, t,\, k)\;\|\mathbf{x}\|_2^2\big)\,, \tag{176}$$

*where, $k \;=\; \|\mathbf{x}\|_\infty^{-2}$ denotes the inverse squared $\ell_\infty$-norm of the input vector, and $\Lambda(\cdot)$ is defined in Notation 1, where $\|X\|_{L^t} \;:=\; \left(\mathbb{E}[|X|^t]\right)^{1/t}$*

*Proof.* According to Equation (229), we have

$$\|X\|_t^t = \|\mathbf{x}\|_\infty^{2t} \sum_{\beta=1}^{t/2} \sum_{\alpha=2\beta}^{t} \frac{D^\beta}{(D\|\mathbf{x}\|_\infty^2)^\alpha} \sum_{V\in([d])^\alpha} |\mathcal{S}_{V,\beta}| \prod_{q\in V} x_q^2.$$

Then, by Claim 17, we get

$$\|X\|_t^t \le \frac{2^{O(t)}}{k^t} \sum_{\beta=1}^{t/2} \sum_{\alpha=2\beta}^{t} \big(M(\alpha,\beta) + N(\alpha,\beta)\big). \tag{235}$$

Assume first that $k \ge Dt$, then by Lemmas 18 and 19, we get

$$\|X\|_t^t \le \frac{2^{O(t)}}{k^t} \sum_{\beta=1}^{t/2} \sum_{\alpha=2\beta}^{t} \big(M(\alpha,\beta) + N(\alpha,\beta)\big), \tag{236}$$

$$\le \frac{2^{O(t)}}{k^t} \left(\frac{k^2 t}{D}\right)^{t/2}, \tag{237}$$

and therefore

$$\|X\|_t = O\left(\sqrt{\frac{t}{D}}\right). \tag{238}$$

Next, assume that $Dt > k \ge \sqrt{Dt}$. Once again, by Lemmas 18 and 19, we get

$$\|X\|_t^t \le \frac{2^{O(t)}}{k^t} \sum_{\beta=1}^{t/2} \sum_{\alpha=2\beta}^{t} \big(M(\alpha,\beta) + N(\alpha,\beta)\big), \tag{239}$$

$$\le \frac{2^{O(t)}}{k^t} \sum_{\beta=1}^{t/2} \sum_{\alpha=2\beta}^{t} \max\left\{ \left(\frac{t}{\ln\frac{2Dt}{k}}\right)^{2t}, \left(\frac{k^2 t}{D}\right)^{t/2} \right\} + \left(\frac{k^2 t}{D}\right)^{t/2}. \tag{17}$$

Thus,

$$\|X\|_t = O\left(\max\left\{ \frac{t}{k\ln\frac{2Dt}{k}}, \sqrt{\frac{t}{D}} \right\}\right). \tag{240}$$

Finally, assume that $\sqrt{Dt} > k$. Again, by Lemmas 18 and 19, we get

$$\|X\|_t^t \leq \frac{2^{O(t)}}{k^t} \sum_{\beta=1}^{t/2} \sum_{\alpha=2\beta}^{t} \big(M(\alpha,\beta) + N(\alpha,\beta)\big), \tag{241}$$

$$\leq \frac{2^{O(t)}}{k^t} \sum_{\beta=1}^{t/2} \sum_{\alpha=2\beta}^{t} \max\left\{ \left(\frac{t}{\ln \frac{2Dt}{k}}\right)^{2t}, \left(\frac{k^2 t}{D}\right)^{t/2} \right\} + \max\left\{ M(\alpha,\beta), \left(\frac{t}{\ln \frac{Dt}{k^2}}\right)^{t} \right\}. \tag{242}$$

Hence,

$$\|X\|_t = O\left( \max\left\{ \frac{t}{k \ln \frac{2Dt}{k}}, \frac{t}{k \ln \frac{Dt}{k^2}}, \sqrt{\frac{t}{D}} \right\} \right), \tag{243}$$

$$\|\mathbf{Cx}\|_2^2 - 1\|_t = O\left( \max\left\{ \frac{t}{k \ln \frac{2Dt}{k}}, \frac{t}{k \ln \frac{Dt}{k^2}}, \sqrt{\frac{t}{D}} \right\} \right), \tag{244}$$

$$\|\mathbf{Cx}\|_2^2 - 1\|_t = O(\Lambda(D,\, t,\, k)). \tag{245}$$

where, $\Lambda(\cdot)$ is a function defined in Notation 1 that depends on the sketch dimension $D$, the moment order $t$, and $k$ denotes inverse squared $\ell_\infty$-norm of the input vector.

□

The next two subsections present the proofs of Approximate Matrix Product and Spectral Approximation. Although the arguments are adapted from Wacker et al. (2023), we provide the full proofs for completeness.

### A.2 Proof of Approximate Matrix Product

*Proof.* Our first objective is to establish the following $L^t$-moment bound:

$$\left\| (\mathbf{Cx})^\top \overline{(\mathbf{Cy})} - \mathbf{x}^\top \mathbf{y} \right\|_{L^t} \;\leq\; \varepsilon\, \delta^{1/t}\, \|\mathbf{x}\|_2\, \|\mathbf{y}\|_2. \tag{246}$$

Without loss of generality, we assume $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$ throughout the proof, since both sides of (246) scale multiplicatively with $\|\mathbf{x}\|_2 \|\mathbf{y}\|_2$.

By Theorem 12, we have

$$\left\| \|\mathbf{Cz}\|_2^2 - \|\mathbf{z}\|_2^2 \right\|_{L^t} \;\leq\; \varepsilon\, \delta^{1/t}\, \|\mathbf{z}\|_2^2, \tag{247}$$

for all $\mathbf{z} \in \mathbb{R}^{d_1 \cdots d_p}$.

The rest of the proof follows (Ahle et al., 2020, Lemma 9). For any vectors $\mathbf{a}, \mathbf{b}$, we recall the identities

$$\|\mathbf{a} - \mathbf{b}\|_2^2 = \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 - 2\, \mathbf{a}^\top \mathbf{b}, \tag{248}$$

$$\|\mathbf{a} + \mathbf{b}\|_2^2 = \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 + 2\, \mathbf{a}^\top \mathbf{b}, \tag{249}$$

which together imply

$$\mathbf{a}^\top \mathbf{b} = \frac{\|\mathbf{a} + \mathbf{b}\|_2^2 - \|\mathbf{a} - \mathbf{b}\|_2^2}{4}. \tag{250}$$

Applying (250) with $\mathbf{a} = \mathbf{C}\mathbf{x}$ and $\mathbf{b} = \mathbf{C}\mathbf{y}$ gives

$$\begin{aligned}
\left\| (\mathbf{C}\mathbf{x})^\top \overline{(\mathbf{C}\mathbf{y})} - \mathbf{x}^\top \mathbf{y} \right\|_{L^t} &= \frac{1}{4} \left\| \|\mathbf{C}(\mathbf{x}+\mathbf{y})\|_2^2 - \|\mathbf{C}(\mathbf{x}-\mathbf{y})\|_2^2 - \|\mathbf{x}+\mathbf{y}\|_2^2 + \|\mathbf{x}-\mathbf{y}\|_2^2 \right\|_{L^t}, \\
&\leq \frac{1}{4} \left( \left\| \|\mathbf{C}(\mathbf{x}+\mathbf{y})\|_2^2 - \|\mathbf{x}+\mathbf{y}\|_2^2 \right\|_{L^t} + \left\| \|\mathbf{C}(\mathbf{x}-\mathbf{y})\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2 \right\|_{L^t} \right), \\
&\leq \frac{\varepsilon \delta^{1/t}}{4} \left( \|\mathbf{x}+\mathbf{y}\|_2^2 + \|\mathbf{x}-\mathbf{y}\|_2^2 \right), \\
&= \frac{\varepsilon \delta^{1/t}}{2} \left( \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 \right) = \varepsilon \delta^{1/t}, \qquad (251)
\end{aligned}$$

where the last equality follows from the normalization $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$. This proves Eqaution (246).

To complete the proof, we convert the moment bound into a tail bound using Markov's inequality, which states that $\Pr(X \geq a) \leq \mathbb{E}[X]/a$ for any $a > 0$.

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d_1 \cdots d_p \times n}$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m) \in \mathbb{R}^{d_1 \cdots d_p \times m}$. Define

$$a = \varepsilon^2 \|\mathbf{X}\|_F^2 \|\mathbf{Y}\|_F^2, \qquad (252)$$

$$X = \left\| (\mathbf{C}\mathbf{X})^\top \overline{(\mathbf{C}\mathbf{Y})} - \mathbf{X}^\top \mathbf{Y} \right\|_F^2. \qquad (253)$$

Then

$$\Pr\left[ \left\| (\mathbf{C}\mathbf{X})^\top \overline{(\mathbf{C}\mathbf{Y})} - \mathbf{X}^\top \mathbf{Y} \right\|_F^2 \geq \varepsilon^2 \|\mathbf{X}\|_F^2 \|\mathbf{Y}\|_F^2 \right] \leq \frac{\varepsilon^2 \delta \sum_{i=1}^n \sum_{j=1}^m \|\mathbf{x}_i\|_2^2 \|\mathbf{y}_j\|_2^2}{\varepsilon^2 \|\mathbf{X}\|_F^2 \|\mathbf{Y}\|_F^2} = \delta. \qquad (254)$$

The claim follows provided that $\mathbf{C}$ has $D$ rows, where $D$ is chosen as in Theorem 12. □

### A.3 Proof of Spectral Approximation of the Gram Matrix

*Proof.* We rephrase Lemma 11 of Ahle et al. (2020) for the case $\lambda > 0$. This ensures that $\mathbf{K} + \lambda \mathbf{I}_n$ is positive definite and that $(\mathbf{K} + \lambda \mathbf{I}_n)^{-1/2}$ exists. The corresponding result for $\lambda = 0$ follows by Fatou's lemma, as in the original proof.

By (Tropp, 2012, Proposition 2.1.1), left- and right-multiplying the spectral inequality (209) by $(\mathbf{K}+\lambda \mathbf{I}_n)^{-1/2}$ preserves positive semidefiniteness. Thus, (209) is equivalent to

$$(1-\varepsilon) I_n \preceq \left( \mathbf{C}\mathbf{A}^{\otimes p} (\mathbf{K} + \lambda I_n)^{-1/2} \right)^\top \overline{\left( \mathbf{C}\mathbf{A}^{\otimes p} (\mathbf{K} + \lambda I_n)^{-1/2} \right)} + \lambda (\mathbf{K} + \lambda I_n)^{-1} \preceq (1+\varepsilon) I_n.$$

Equivalently,

$$\left\| \left( \mathbf{C}\mathbf{A}^{\otimes p} (\mathbf{K} + \lambda I_n)^{-1/2} \right)^\top \overline{\left( \mathbf{C}\mathbf{A}^{\otimes p} (\mathbf{K} + \lambda I_n)^{-1/2} \right)} + \lambda (\mathbf{K} + \lambda I_n)^{-1} - I_n \right\|_2 \leq \varepsilon.$$

Now define

$$\mathbf{Z} := \mathbf{A}^{\otimes p} (\mathbf{K} + \lambda I_n)^{-1/2}, \qquad (255)$$

So that

$$\begin{aligned}
\mathbf{Z}^\top \mathbf{Z} &= (\mathbf{K} + \lambda I_n)^{-1/2} \mathbf{K} (\mathbf{K} + \lambda I_n)^{-1/2}, \qquad (256) \\
&= (\mathbf{K} + \lambda I_n)^{-1} (\mathbf{K} + \lambda I_n - \lambda I_n)(\mathbf{K} + \lambda I_n)^{-1}, \qquad (257) \\
&= I_n - \lambda (\mathbf{K} + \lambda I_n)^{-1}. \qquad (258)
\end{aligned}$$

Substituting into (255) yields

$$\left\|(\mathbf{CZ})^\top\overline{(\mathbf{CZ})} - \mathbf{Z}^\top\mathbf{Z}\right\|_2 \;\le\; \varepsilon. \tag{259}$$

We now apply the Frobenius-norm error bound (Theorem 12), which holds for all matrices. Setting $\mathbf{X} = \mathbf{Y} = \mathbf{Z} \in \mathbb{C}^{d^p \times n}$, we obtain

$$\Pr\left[\left\|(\mathbf{CZ})^\top\overline{(\mathbf{CZ})} - \mathbf{Z}^\top\mathbf{Z}\right\|_2 \ge \varepsilon\right] \le \Pr\left[\left\|(\mathbf{CZ})^\top\overline{(\mathbf{CZ})} - \mathbf{Z}^\top\mathbf{Z}\right\|_F \ge \varepsilon\right] \le \delta, \tag{260}$$

provided that $\mathbf{C}$ has $D\,\|\mathbf{Z}\|_F^2$ rows. Finally, we compute

$$\|\mathbf{Z}\|_F^2 = \operatorname{tr}(\mathbf{Z}^\top\mathbf{Z}) = \operatorname{tr}\big(I_n - \lambda(\mathbf{K} + \lambda I_n)^{-1}\big)\,, \tag{261}$$

$$= \sum_{i=1}^{n} \frac{\lambda_i(\mathbf{K})}{\lambda_i(\mathbf{K}) + \lambda} = \operatorname{tr}\big(\mathbf{K}(\mathbf{K} + \lambda I_n)^{-1}\big) = s_\lambda(\mathbf{K}), \tag{262}$$

where $\{\lambda_i(\mathbf{K})\}_{i=1}^n$ denote the eigenvalues of $\mathbf{K}$ and $0 \le s_\lambda(\mathbf{K}) \le n$ is the $\lambda$-statistical dimension. Substituting into (260) completes the proof. □