Improving TensorSketch Using Complex Random Variables

Amit Sharma[*1] **Mohammad Azhar Khan**[*1] **Rameshwar Pratap**[1] **Keegan Kang**[2]

[1]Department of Computer Science and Engineering, IIT Hyderabad, India
[2]Department of Mathematics and Statistics, Bucknell University, USA

Abstract

`TensorSketch` by Pham and Pagh [2013], Kar and Karnick [2012] provides efficient sketching algorithms for high-dimensional polynomial kernels $\mathbf{x}^{\otimes p} \in \mathbb{R}^{d^p}$. Kar and Karnick [2012] uses dense Johnson-Lindenstrauss (JL)-type projections with computational cost $O(pDd)$, where D denotes the sketch dimension, whereas Pham and Pagh [2013] extends the sparse `CountSketch` [Charikar et al., 2004] algorithm, yielding a faster algorithm for high-dimensional sparse inputs with running time $O\big(p(\mathrm{nnz}\,(\mathbf{x}) + D\log D)\big)$. However, the variance of both estimators grows exponentially with the polynomial degree p, scaling as $3^p/D$. Recent work by Wacker et al. [2023] showed that using complex-valued distribution reduces this dependence to $2^p/D$ for the approach of Kar and Karnick [2012]. However, their method relies on dense JL-type projections with computational cost $O(pDd)$ and does not extend to the algorithm of Pham and Pagh [2013].

In this work, we introduce a simple variant of `TensorSketch` [Pham and Pagh, 2013] that achieves the same variance bound as Wacker et al. [2023], while retaining its advantage of the input-sparsity running time. We validate our results with supporting experiments on synthetic and real-world datasets.

1 INTRODUCTION

Polynomial kernels (Definition 3) are widely used in machine learning to model higher-order, non-linear interactions among input features. For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, a degree-$p$ polynomial kernel is defined as $k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y}\rangle^p$. This kernel is equivalent to mapping $\mathbf{x} \in \mathbb{R}^d$ to its p-fold Kronecker product $\mathbf{x}^{\otimes p} \in \mathbb{R}^{d^p}$, which is $\langle \mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p}\rangle \equiv \langle \mathbf{x}, \mathbf{y}\rangle^p$.

As the dimension d^p grows exponentially with p, directly computing these inner products is computationally infeasible. A dense Johnson-Lindenstrauss (JL) transform [Johnson and Lindenstrauss, 1984] can reduce their dimensionality while approximately preserving their pairwise inner product, however applying it requires time proportional to d^p, which is exponential in p.

To address this inefficiency, prior works such as Kar and Karnick [2012] and Pham and Pagh [2013] proposed randomized sketching techniques to approximate polynomial kernels efficiently. Kar and Karnick [2012] introduced random feature maps based on JL-type projections. They define a randomized linear map $\mathbf{S} : \mathbb{R}^{d^p} \to \mathbb{R}^D$ by $\mathbf{S}(\mathbf{x}^{\otimes p}) := (\mathbf{W}_1\mathbf{x} \odot \cdots \odot \mathbf{W}_p\mathbf{x})/\sqrt{D}$, where each $\mathbf{W}_i \in \mathbb{R}^{D\times d}$ is a random projection matrix (e.g., Gaussian or Rademacher with *i.i.d.* entries), and $\odot$ denotes the element-wise (Hadamard) product. Such that $\widehat{k}\,(\mathbf{x}, \mathbf{y}) := \langle \mathbf{S}(\mathbf{x}^{\otimes p}), \mathbf{S}(\mathbf{y}^{\otimes p})\rangle$. This estimator can be termed as a JL-type variant of `TensorSketch`, and its computational cost scales with $O(pDd)$. Pham and Pagh [2013] propose `TensorSketch`, which combines `CountSketch` [Charikar et al., 2004] with Fast Fourier Transform (FFT)-based convolution to compute the sketch implicitly. This avoids forming the full vector and runs in input-sparsity time $O\big(p(\mathrm{nnz}\,(\mathbf{x}) + D\log D)\big)$, making it preferred over Kar and Karnick [2012] for high-dimensional sparse data. The variance of both algorithms Kar and Karnick [2012], Pham and Pagh [2013, 2025] grows as $3^p/D$.

Recent progress by Wacker et al. [2024] introduced a complex-valued variant of JL-type `TensorSketch`. It improves the variance dependence of the sketch on the polynomial degree, reducing it from 3^p to 2^p. This improvement is achieved by incorporating complex-valued randomness into the sketch. However, the resulting sketch vectors are complex-valued, which cannot be directly compared with its real-valued counterpart. To address this, Wacker et al. [2023] proposed the *Complex-to-Real (CtR)* construction. It first

[*] Equal contribution.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:6178–6203.

computes a complex random feature map of the embedding dimension half of the size of its real counterpart and then forms a real embedding by concatenating its real and imaginary parts. This preserves inner products and achieves improved variance bounds while yielding a real-valued sketch. However, as a JL-type method, it still requires dense multiplications with cost $O(pDd)$, which can be inefficient for high-dimensional sparse data.

In this work, we address the above limitation by developing the *Complex-to-Real (CtR)* variant of `TensorSketch`. The proposed constructions obtain variance improvements comparable to the complex JL-type estimator of Wacker et al. [2023], while retaining the input-sparsity running time guarantees of `TensorSketch` [Pham and Pagh, 2013].

Contributions. We propose a *Complex-to-Real* variant of `TensorSketch` for degree-p polynomial kernels. It combines a random function whose values are drawn independently and uniformly from the fourth roots of unity with FFT-based tensor sketching while producing real-valued embeddings. We prove that the resulting estimator is unbiased for $\langle \mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p} \rangle$. We also derive an upper bound on variance with 2^p-type dependence on the degree, improving over the classical `TensorSketch` [Pham and Pagh, 2013] variance scaling, while retaining the sketching time $O(p(\mathrm{nnz}\,(\mathbf{x}) + D \log D))$. The algorithm is defined in Definition 8, and its theoretical guarantees are stated in Theorem 2.

We note that the use of complex-valued random variables in sketching algorithms has been explored in prior work. For example, Wacker et al. [2023, 2024] employ complex random variables to construct sketches for polynomial kernels, while Meyer and Avron [2026] uses them for trace estimation of implicit matrices. The key idea in Wacker et al. [2023, 2024] is to sketch $\mathbf{x}^{\otimes p}$ using element-wise products of independent sketches of the factors $(\mathbf{x}_1, \ldots, \mathbf{x}_p)$. Owing to this independence structure, Khintchine's inequality (Definition 4) can be applied, and a simple induction on p yields a tighter upper bound in the complex setting than in the real-valued case.

In contrast, `TensorSketch` is built upon the hash-based `CountSketch` framework, where the sketch components are not independent and no direct analogue of Khintchine's inequality is available. As a result, the techniques from prior work do not extend to our setting. Moreover, our analysis of `CountSketch` with complex-valued random variables (Theorem 5, Appendix) shows that, by itself, the complex construction of `CountSketch` provides no variance reduction over its real-valued version. Therefore, our result demonstrating improved variance for `TensorSketch` with complex random variables is non-trivial and stems from the vanishing of certain cross terms in the variance analysis, which does not occur in the real-valued setting.

Implications of our results. `TensorSketch` [Pham and Pagh, 2013] enables efficient training of linear SVMs [Sun et al., 2018, Li et al., 2019], supports scalable deep learning and the theoretical analysis of over-parameterized neural networks [Yehudai and Shamir, 2019, Zandieh et al., 2021], and has been successfully applied to compact bilinear pooling for fine-grained visual recognition [Gao et al., 2016] as well as multimodal fusion architectures [Fukui et al., 2016]. Polynomial kernels themselves are widely adopted in applications including natural language processing [Goldberg and Elhadad, 2008], recommender systems [Rendle, 2010], and genomics [Aschard, 2016]. Our proposed CtR `TensorSketch` achieves more accurate estimates than `TensorSketch` [Pham and Pagh, 2013] while maintaining the same asymptotic time complexity, making it a promising alternative for the above applications.

Organization of the paper. The remainder of this paper is organized as follows. Section 2 reviews prior work on polynomial kernel approximation and randomized sketching. Section 3 presents necessary background, including `CountSketch`, `TensorSketch`, polynomial kernels, and the CtR framework. Section 4, proposes our algorithm and states theoretical guarantees on its accuracy and efficiency. Section 5 reports empirical comparisons of real and JL-type methods with our approach. Section 6 concludes and outlines some future research directions.

2 RELATED WORK

Polynomial kernels can be expressed via tensor feature maps, where the representation is 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 interactions. Explicit construction requires $O(\prod_{i=1}^{p} d_i)$ memory and is infeasible even for moderate p or dimensions d. To avoid this blow-up, prior work proposes sketching methods that compute $\mathbf{S}(\bigotimes_{i=1}^{p} \mathbf{x}_i)$ without forming the full tensor, including JL-type product embeddings [Kar and Karnick, 2012] and hashing-based `TensorSketch` constructions built on `CountSketch` [Pham and Pagh, 2013, 2025]. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, both `TensorSketch` estimators have variance $\leq \frac{3^p-1}{D}\left(\|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2\right)$, which yields 3^p-type variance growth. Both JL-type and hashing-based estimators scale as $\Theta(3^p/D)$ for degree-p features, but `TensorSketch` [Pham and Pagh, 2013, 2025] is more efficient due its input-sparsity time complexity.

Recent work shows that complex-valued distributions reduces the variance of randomized feature maps. Complex JL-type constructions for polynomial kernels achieve lower variance than their real-valued counterparts [Wacker et al., 2024]. In particular, replacing real Rademacher or Gaussian variables with complex-valued distributions improves the variance dependence from 3^p-type to 2^p-type. Further work by Wacker et al. [2023] introduced the *Complex-to-Real*

(CtR) construction, preserving the variance improvement of complex-valued sketches while yielding real-valued embeddings. These embeddings enables direct comparison with real sketches. However, these approaches rely on dense JL-style projections and therefore incur higher computational cost especially for high-dimensional sparse data.

Building on the hashing-based sketching framework of [Pham and Pagh, 2013], we introduce a *Complex-to-Real (CtR)* variant of `TensorSketch` for polynomial kernel approximation. Our method uses fourth roots-of-unity instead of real Rademacher variables and applies a structured *complex-to-real* transformation, yielding unbiased real-valued embeddings with improved variance dependence.

In contrast, our method achieves variance reduction through a lightweight modification of the sketch construction. This is followed by a *Complex-to-Real (CtR)* conversion to produce a real-valued embedding that makes it possible to compare it fairly with its real counterpart. The resulting sketch preserves the original structure and input-sparsity running time while achieving provable variance improvements over original `TensorSketch` [Pham and Pagh, 2013].

Variance reduction for randomized sketching has been widely studied using statistical variance reduction techniques such as control variates (CV) and maximum likelihood estimation (MLE). The CV method reduces variance by leveraging a correlated auxiliary variable with a known expectation, while MLE estimates unknown parameters by maximizing the likelihood of the observed data, often yielding statistically efficient estimators. These techniques have been extensively investigated for a variety of classical randomized sketching methods, including Johnson–Lindenstrauss (JL) transforms Li et al. [2006], CountSketch Pratap and Kulkarni [2021], Tug-of-War sketches Pratap et al. [2021], signed random projections Kang and Wong [2018], feature hashing Verma et al. [2022], and compressed matrix multiplication Verma et al. [2025], among others. However, to the best of our knowledge, analogous variance reduction techniques have not yet been developed for `TensorSketch`. While these approaches can substantially reduce estimator variance, they typically incur additional computational and algorithmic overhead.

3 PRELIMINARIES

Notation: We use the following notation in the paper. Bold lowercase letters (e.g., $\mathbf{x}, \mathbf{y}$) denote vectors, bold uppercase letters (e.g., $\mathbf{S}, \mathbf{T}$) denote matrices. For a positive integer D, we write $[D] := \{1, 2, \ldots, D\}$ for the corresponding index set. For $d, p \in \mathbb{N}$, we denote by $[d]^p$ the set of all p-tuples $(i_1, \ldots, i_p)$ with $i_j \in [d]$. The sets $\mathbb{R}$, $\mathbb{Z}$ and $\mathbb{C}$ denote the real, integer and complex domains, respectively. 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 $\|\mathbf{S}\|_F$ 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 $\odot$ denotes the elementwise (Hadamard) product. The indicator function is denoted by $\mathbf{1}\,[\cdot]$. The probabilistic quantities use $\Pr[\cdot]$ for probability, $\mathbb{E}\,[\cdot]$ for expectation and $\mathrm{Var}\,(\cdot)$ for variance.

Definition 1 (`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 2-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, $z_j = \sum_{h(i)=j} s(i)\, y_i,\ \forall j = \{1, \ldots, D\}$. The time complexity of computing the `CountSketch` is $O(\mathrm{nnz}\,(\mathbf{y}))$.

For pairwise inner–product analysis, let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and define the `CountSketch` inner–product estimator as

$$\widehat{k}\,(\mathbf{x}, \mathbf{y}) := \langle \mathbf{T}\mathbf{x}, \mathbf{T}\mathbf{y} \rangle.$$

Then the estimator is unbiased

$$\mathbb{E}\left[\widehat{k}\,(\mathbf{x}, \mathbf{y})\right] = \langle \mathbf{x}, \mathbf{y} \rangle,$$

and its variance satisfies

$$\mathrm{Var}\left(\widehat{k}\,(\mathbf{x}, \mathbf{y})\right) \leq \frac{1}{D}\Big(\langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|_2^2\, \|\mathbf{y}\|_2^2 - 2\sum_i x_i^2 y_i^2\Big).$$

Definition 2 (`TensorSketch` of Degree p [Pham and Pagh, 2013, 2025])**.** *Let $p \geq 2$. For each $t \in [p]$, let $h_t : [d] \to [D]$ be 2-wise independent hash functions and $\sigma_t : [d] \to \{-1, +1\}$ be 4-wise independent random sign functions. The degree-p `TensorSketch` matrix $\mathbf{S} \in \mathbb{R}^{D \times d^p}$ is defined for all $r \in [D]$ and $(i_1, \ldots, i_p) \in [d]^p$ by*

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

For any $\mathbf{x} \in \mathbb{R}^d$, the sketch $\mathbf{S}(\mathbf{x}^{\otimes p})$ can be computed implicitly in time $O(p(\mathrm{nnz}\,(\mathbf{x}) + D \log D))$ using FFT-based convolution, without explicitly forming $\mathbf{x}^{\otimes p}$.
Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, then estimate for pairwise inner–product is defined as follows:

$$\widehat{k}\,(\mathbf{x}, \mathbf{y}) := \langle \mathbf{S}\big(\mathbf{x}^{\otimes p}\big), \mathbf{S}\big(\mathbf{y}^{\otimes p}\big) \rangle.$$

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

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

and its variance satisfies

$$\mathrm{Var}\left(\widehat{k}\left(\mathbf{x},\mathbf{y}\right)\right) \leq \frac{1}{D}\Bigg(\Big(2\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2 \cdots \\ \cdots - 2\sum_{i=1}^{d} x_i^2 y_i^2\Big)^p - \langle \mathbf{x},\mathbf{y}\rangle^{2p}\Bigg), \\ \leq \frac{3^p-1}{D}\,\|\mathbf{x}\|_2^{2p}\,\|\mathbf{y}\|_2^{2p}.$$

Definition 3 (Polynomial Kernel [Scholkopf and Smola, 2001])**.** *We consider polynomial kernels of degree $p \in \mathbb{N}$ of the form*

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

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

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

With this transformation, the kernel admits a homogeneous representation,

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

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 4 (Khintchine Inequality Haagerup and Musat [2007], Wacker et al. [2023])**.** *Let $\mathbf{x} = (x_1,\ldots,x_d) \in \mathbb{R}^d$ and let $\{\varepsilon_i\}_{i=1}^d$ be independent Rademacher random variables. For every $p > 0$, there exists a constant K_p such that*

$$\left(\mathbb{E}\left|\sum_{i=1}^{d} x_i\varepsilon_i\right|^p\right)^{1/p} \leq K_p\|\mathbf{x}\|_2. \tag{1}$$

The optimal constants are known explicitly.

- *If $\varepsilon_i \in \{-1,+1\}$ are real-valued Rademacher variables, then*

$$K_p = \begin{cases} 1, & 0 < p \leq 2, \\ \sqrt{2}\,\pi^{-\frac{1}{2p}}\Gamma\left(\frac{p+1}{2}\right)^{\frac{1}{p}}, & p > 2. \end{cases} \tag{2}$$

- *If ε_i are complex Rademacher variables taking values in $\{1,-1,i,-i\}$ uniformly at random, then*

$$\hat{K}_p = \Gamma\left(\frac{p}{2}+1\right)^{\frac{1}{p}}. \tag{3}$$

Moreover, for every $p > 2$, the complex constant $\hat{K}_p$ is strictly smaller than the corresponding real constant K_p.

Definition 5 (Framework for Complex-to-Real (CtR) Sketches)**.** *Wacker et al. [2023] suggests a framework of analysing sketching algorithms involving complex random variables. Let $z = a + ib$ be a complex random variable with $a,b \in \mathbb{R}$, we have $|z|^2 = a^2 + b^2$ and $\mathrm{Re}\left\{z^2\right\} = a^2 - b^2$. Combining both gives $a^2 = \frac{1}{2}\left(|z|^2 + \mathrm{Re}\left\{z^2\right\}\right)$. The scalar a is real-valued and its variance $\mathrm{Var}\,(a) = \mathbb{E}\left[a^2\right] - \mathbb{E}\left[a\right]^2$ is therefore*

$$\mathrm{Var}\,(a) = \frac{1}{2}\mathrm{Re}\{\mathbb{E}[|z|^2] + \mathbb{E}[z^2] + 2\mathbb{E}[a]^2\}. \tag{4}$$

Let $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$ and let $\Phi_{\mathrm{C}} : \mathbb{R}^{d^p} \to \mathbb{C}^{\frac{D}{2}}$. Then $\widehat{k}_{\mathrm{C}}\left(\mathbf{x},\mathbf{y}\right) := \Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p})\overline{\Phi_{\mathrm{C}}(\mathbf{y}^{\otimes p})}^\top \in \mathbb{C}$ is a complex-valued estimate of the polynomial kernel $k(\mathbf{x},\mathbf{y})$, and hence $\widehat{k}_{\mathrm{C}}\left(\mathbf{x},\mathbf{y}\right) = \mathrm{Re}\{\widehat{k}_{\mathrm{C}}\left(\mathbf{x},\mathbf{y}\right)\} + i\mathrm{Im}\{\widehat{k}_{\mathrm{C}}\left(\mathbf{x},\mathbf{y}\right)\}$. The Complex-to-Real *sketch is defined as follows:*

$$\begin{aligned} &\widehat{k}_{\mathrm{CtR}}\left(\mathbf{x},\mathbf{y}\right) := \mathrm{Re}\{\widehat{k}_{\mathrm{C}}\left(\mathbf{x},\mathbf{y}\right)\}, \\ &= \mathrm{Re}\left\{\Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p})\right\}^\top \mathrm{Re}\left\{\Phi_{\mathrm{C}}(\mathbf{y}^{\otimes p})\right\} + \\ &\quad \cdots + \mathrm{Im}\left\{\Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p})\right\}^\top \mathrm{Im}\left\{\Phi_{\mathrm{C}}(\mathbf{y}^{\otimes p})\right\}, \\ &= \Phi_{\mathrm{CtR}}(\mathbf{x})^\top\Phi_{\mathrm{CtR}}(\mathbf{y}). \end{aligned}$$

where, $\Phi_{\mathrm{CtR}}(\mathbf{x}) := \left[\mathrm{Re}\left\{\Phi_{\mathrm{C}}\right\}, \mathrm{Im}\left\{\Phi_{\mathrm{C}}\right\}\right] \in \mathbb{R}^D$. We now derive the variance of $\widehat{k}_{\mathrm{CtR}}\left(\mathbf{x},\mathbf{y}\right)$ using Equation (4)*:*

$$\begin{aligned} &\mathrm{Var}\left(\widehat{k}_{\mathrm{CtR}}\left(\mathbf{x},\mathbf{y}\right)\right) \\ &= \frac{1}{2}\mathrm{Re}\{\mathbb{E}[|\widehat{k}_C|^2] + \mathbb{E}[\widehat{k}_C^2] + 2\mathbb{E}[\mathrm{Re}\{\widehat{k}_C\}]^2\} \\ &= \frac{1}{2}\mathrm{Re}\{\mathbb{E}[|\widehat{k}_C|^2] + \mathbb{E}[\widehat{k}_C^2] + 2\mathbb{E}[\widehat{k}_C]^2\}, \end{aligned} \tag{5}$$

where, $\widehat{k}_C := \widehat{k}_{\mathrm{C}}\left(\mathbf{x},\mathbf{y}\right) \in \mathbb{C}$. This framework plays a key role in computing the expectation and variance of $\widehat{k}_{\mathrm{CtR}}\left(\mathbf{x},\mathbf{y}\right)$.

Definition 6 (*Complex-to-Real* Polynomial Sketch [Wacker et al., 2023])**.** *Let $p \in \mathbb{N}$ and $D = 2k$ for some $k \in \mathbb{N}$. For each $i \in [p]$, let $\mathbf{W}_i \in \mathbb{C}^{D/2\times d}$ be a random matrix whose rows are sampled independently from a zero-mean distribution satisfying $\mathbb{E}\left[\mathbf{w}\mathbf{w}^*\right] = \mathbf{I}_d$ (e.g., complex Gaussian or complex Rademacher) where $*$ refers to conjugate-transpose. Define the complex random feature map*

$$\Phi_C(\mathbf{x}) := \sqrt{\frac{2}{D}}\left(\mathbf{W_1}\mathbf{x} \odot \mathbf{W_2}\mathbf{x} \odot \cdots \odot \mathbf{W_p}\mathbf{x}\right) \in \mathbb{C}^{D/2},$$

Then, the Complex-to-Real (CtR) sketch *of $\mathbf{x} \in \mathbb{R}^d$ is the real-valued vector $\Phi_{\mathrm{CtR}}(\mathbf{x}) \in \mathbb{R}^D$ defined by*

$$\Phi_{\mathrm{CtR}}(\mathbf{x}) := \begin{pmatrix} \mathrm{Re}\left\{\Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p})\right\} \\ \mathrm{Im}\left\{\Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p})\right\} \end{pmatrix} \in \mathbb{R}^D.$$

For any $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$, the kernel estimator can be defined as

$$\widehat{k}_{\mathrm{CtR}}\left(\mathbf{x},\mathbf{y}\right) = \Phi_{\mathrm{CtR}}(\mathbf{x})^\top\Phi_{\mathrm{CtR}}(\mathbf{y}).$$

For both Gaussian and Rademacher constructions, the CtR estimator satisfied the following:

$$\mathbb{E}\left[\widehat{k}_{\mathrm{CtR}}(\mathbf{x},\mathbf{y})\right] = \langle \mathbf{x},\mathbf{y}\rangle^p,$$
$$\mathrm{Var}\left(\widehat{k}_{\mathrm{CtR}}(\mathbf{x},\mathbf{y})\right) \leq \frac{2^{p+1}-2}{D}\|\mathbf{x}\|_2^{2p}\|\mathbf{y}\|_2^{2p}.$$

4 *COMPLEX-TO-REAL(CTR)* `TENSORSKETCH`

In this section, we first define Complex `CountSketch` (Definition 7). Building on its circular convolution structure, we then introduce *Complex-to-Real* `TensorSketch`. (Definition 8) along with the associated kernel. Theorem 2 discusses the guarantees of the estimators proposed in Definition 8.

Definition 7 (Complex `CountSketch` Mapping)**.** *Let $\mathbf{x} \in \mathbb{R}^d$ and D be the sketching dimension. Sample 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 j-th coordinate of the sketched vector $\mathbf{Cx}$ is given by

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

We define CtR `TensorSketch`, which extends the Complex `CountSketch` by applying p independent sketches and combining them using FFT-based convolution to efficiently sketch the vector $\mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p} \in \mathbb{R}^{d^p}$.

Definition 8. *(*Complex-to-Real (CtR) `TensorSketch` *Mapping) Let $\mathbf{x} \in \mathbb{R}^d$ and fix an integer $p \geq 1$. Define a Complex `TensorSketch` mapping $\mathbf{C} : \mathbb{R}^{d^p} \to \mathbb{C}^{D/2}$ 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 $\mathbf{C}_r \in \mathbb{C}^{D/2\times d}$ denote the Complex `CountSketch` matrix induced by functions (h_r, s_r) as in Definition 7. 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

$$\Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p}) := \mathbf{C}\mathbf{x}^{\otimes p}, \tag{6}$$
$$= \mathrm{FFT}^{-1}\left(\prod_{r=1}^{p}\mathrm{FFT}(\mathbf{C}_r\mathbf{x})\right) \in \mathbb{C}^{D/2}, \tag{7}$$

where the product is taken element-wise. We define the CtR sketch as

$$\Phi_{\mathrm{CtR}}(\mathbf{x}^{\otimes p}) := \Big(\mathrm{Re}\{\Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p})_1\}, \ldots, \mathrm{Re}\{\Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p})_{D/2}\},$$
$$\mathrm{Im}\{\Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p})_1\}, \ldots, \mathrm{Im}\{\Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p})_{D/2}\}\Big)^{\top} \in \mathbb{R}^D.$$

A detailed proof of results (Lemma 1, 4, and Theorem 2) stated in this section is presented in Appendix A.1.

Lemma 1. *Let $\omega = e^{2\pi i/4} = i$, and let $s : [d] \to \{1, \omega, \omega^2, \omega^3\}$ be a random function such that the values $\{s(i)\}_{i\in[d]}$ are drawn independently and uniformly from the four fourth roots of unity. Then, for every $i \in [d]$, the following identities hold*

$$\mathbb{E}\left[s(i)\right] = 0,$$
$$\mathbb{E}[|s(i)|^2] = 1,$$
$$\mathbb{E}\left[s(i)^2\right] = 0.$$

Furthermore, for any pair of distinct indices $i \neq j$, independence implies

$$\mathbb{E}[s(i)\,\overline{s(j)}] = \mathbb{E}[s(i)]\,\mathbb{E}\left[\overline{s(j)}\right] = 0.$$

Theorem 2 establishes the unbiasedness, variance bound, and sketching time of the CtR `TensorSketch`.

Theorem 2. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p} \in \mathbb{R}^{d^p}$. Let Φ_{CtR} denote a* Complex-to-Real `TensorSketch` *as stated in Definition 8. Let $\widehat{k}_{\mathrm{CtR}}(\mathbf{x},\mathbf{y}) := \Phi_{\mathrm{CtR}}(\mathbf{x}^{\otimes p})^{\top}\Phi_{\mathrm{CtR}}(\mathbf{y}^{\otimes p})$, then $\widehat{k}_{\mathrm{CtR}}(\mathbf{x},\mathbf{y})$ satisfies*

$$\mathbb{E}\left[\widehat{k}_{\mathrm{CtR}}(\mathbf{x},\mathbf{y})\right] = \langle \mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p}\rangle = \langle \mathbf{x},\mathbf{y}\rangle^p,$$
$$\mathrm{Var}\left(\widehat{k}_{\mathrm{CtR}}(\mathbf{x},\mathbf{y})\right) \leq \frac{1}{D}\Bigg(\Bigg(\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - \cdots$$
$$\cdots - \sum_{i=1}^{d} x_i^2 y_i^2\Bigg)^p - \langle \mathbf{x},\mathbf{y}\rangle^{2p}\Bigg),$$
$$\leq \frac{2^p-1}{D}\,\|\mathbf{x}\|_2^{2p}\,\|\mathbf{y}\|_2^{2p}.$$

Moreover, the sketches $\Phi_{\mathrm{CtR}}(\mathbf{x}^{\otimes p})$ and $\Phi_{\mathrm{CtR}}(\mathbf{y}^{\otimes p})$ 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.

Algorithm	Variance	Sketching Time
`TensorSketch`(CtR)	$\frac{2^{p+1}-2}{D}\ \|\mathbf{x}\|_2^{2p}\ \|\mathbf{y}\|_2^{2p}$	$O(p(\mathrm{nnz}\,(\mathbf{x})+\mathrm{nnz}\,(\mathbf{y})+D\log D))$
`TensorSketch`(Real)[Pham and Pagh, 2013, 2025]	$\frac{3^{p}-1}{D}\ \|\mathbf{x}\|_2^{2p}\ \|\mathbf{y}\|_2^{2p}$	$O(p(\mathrm{nnz}\,(\mathbf{x})+\mathrm{nnz}\,(\mathbf{y})+D\log D))$
JL(CtR Radamacher)[Wacker et al., 2023]	$\frac{2^{p+1}-2}{D}\ \|\mathbf{x}\|_2^{2p}\ \|\mathbf{y}\|_2^{2p}$	$O(pDd)$
JL(CtR Guassian)[Wacker et al., 2023]	$\frac{2^{p+1}-2}{D}\ \|\mathbf{x}\|_2^{2p}\ \|\mathbf{y}\|_2^{2p}$	$O(pDd)$

Table 1: **Comparison of variance bounds and sketching time for CtR `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. All methods provide unbiased estimates of the inner product between the vectors $\mathbf{x}^{\otimes p},\mathbf{y}^{\otimes p}\in\mathbb{R}^{d^p}$, i.e., $\langle\mathbf{x}^{\otimes p},\mathbf{y}^{\otimes p}\rangle$.

Proof. According to Definition 5, we have $\widehat{k}_{\mathrm{CtR}}(\mathbf{x},\mathbf{y}) := \mathrm{Re}\left\{\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})\right\}$, and therefore we first analyze $\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})$. Recall that $\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y}) := \Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p})\overline{\Phi_{\mathrm{C}}(\mathbf{y}^{\otimes p})}^{\top}$, where $\Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p}) := \mathbf{C}\mathbf{x}^{\otimes p}$ and $\Phi_{\mathrm{C}}(\mathbf{y}^{\otimes p}) := \mathbf{C}\mathbf{y}^{\otimes p}$ according to Definition 8. In particular, $\mathbf{C}\mathbf{x}^{\otimes p},\mathbf{C}\mathbf{y}^{\otimes p}\in\mathbb{C}^{D/2}$ denote the complex `CountSketch` of the vectors $\mathbf{x}^{\otimes p},\mathbf{y}^{\otimes p}\in\mathbb{R}^{d^p}$, constructed using the derived functions $H:[d]^p\mapsto[D/2]$ and $S:[d]^p\to\{1,\omega,\omega^2,\omega^3\}$ defined as

$$H(i_1,\ldots,i_p)=\left(\sum_{j=1}^{p}h_j(i_j)\right)\bmod D/2,$$
$$S(i_1,\ldots,i_p)=\prod_{j=1}^{p}s_j(i_j).$$

In the proof, we denote $X:=\mathbf{x}^{\otimes p}$ and $Y:=\mathbf{y}^{\otimes p}$. Let $u,v\in[d]^p$ be the multi-indices corresponding to the entries of $X,Y\in\mathbb{R}^{d^p}$. For a multi-index $u=(j_1,\ldots,j_p)$, the entry of X is defined as $X_u=\prod_{k=1}^{p}x_{j_k}$, where the associated linearized index is $u=1+\sum_{k=1}^{p}(j_k-1)\prod_{k'=1}^{p-k}d$. The entries of Y are defined analogously.

We begin by analyzing $\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})$ as defined below,

$$\begin{aligned}\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y}) &:= \Phi_{\mathrm{C}}(X)\overline{\Phi_{\mathrm{C}}(Y)}^{\top}=\langle\mathbf{C}X,\overline{\mathbf{C}Y}\rangle,\\ &=\sum_{u,v\in[d]^p}X_u\,Y_v\,S(u)\,\overline{S(v)}\,\mathbf{1}\left[H(u)=H(v)\right],\\ &=\langle X,Y\rangle+\sum_{u\neq v}X_u\,Y_v\,S(u)\,\overline{S(v)}\,\mathbf{1}\left[H(u)=H(v)\right].\end{aligned}$$

Using Lemma 1, we have $\mathbb{E}\left[S(u)\,\overline{S(v)}\right]=0$ for all $u\neq v$. Hence,

$$\mathbb{E}\left[\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})\right]=\langle X,Y\rangle=\langle\mathbf{x}^{\otimes p},\mathbf{y}^{\otimes p}\rangle=\langle\mathbf{x},\mathbf{y}\rangle^p.$$

According to Definition 5, we know $\widehat{k}_{\mathrm{CtR}}(\mathbf{x},\mathbf{y}) := \mathrm{Re}\left\{\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})\right\}$, we have

$$\mathbb{E}\left[\widehat{k}_{\mathrm{CtR}}(\mathbf{x},\mathbf{y})\right]=\mathbb{E}\left[\mathrm{Re}\left\{\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})\right\}\right]=\langle\mathbf{x},\mathbf{y}\rangle^p. \quad (8)$$

Using Equation 5, the variance of $\widehat{k}_{\mathrm{CtR}}(\mathbf{x},\mathbf{y})$ is

$$\mathrm{Var}\left(\widehat{k}_{\mathrm{CtR}}(\mathbf{x},\mathbf{y})\right)=\frac{1}{2}\,\mathrm{Re}\Bigg\{\mathbb{E}\left[\left|\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})\right|^2\right]+\cdots$$
$$\cdots+\mathbb{E}\left[\left(\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})\right)^2\right]-2\left(\mathbb{E}\left[\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})\right]\right)^2\Bigg\}. \quad (9)$$

For the variance analysis, we need to compute $\mathbb{E}\left[\left|\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})\right|^2\right]$ and $\mathbb{E}\left[\left(\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})\right)^2\right]$. By applying Lemma 1 and Lemma 4, we obtain the following bounds:

$$\mathbb{E}\left[\left|\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})\right|^2\right]\le\langle\mathbf{x},\mathbf{y}\rangle^{2p}+\frac{2}{D}\Bigg(\Bigg(\langle\mathbf{x},\mathbf{y}\rangle^2+$$
$$\cdots+\|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2-\sum_{i=1}^{d}x_i^2y_i^2\Bigg)^p-\langle\mathbf{x},\mathbf{y}\rangle^{2p}\Bigg), \quad (10)$$

and

$$\mathbb{E}\left[\left(\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})\right)^2\right]\le\langle\mathbf{x},\mathbf{y}\rangle^{2p}+\frac{2}{D}\Bigg(\Bigg(2\langle\mathbf{x},\mathbf{y}\rangle^2-$$
$$\cdots-\sum_{i=1}^{d}x_i^2y_i^2\Bigg)^p-\langle\mathbf{x},\mathbf{y}\rangle^{2p}\Bigg). \quad (11)$$

Substituting Equations (10), (11), and (8) into Equation (9) yields

$$\mathrm{Var}\left(\widehat{k}_{\mathrm{CtR}}(\mathbf{x},\mathbf{y})\right)\le\frac{2^{p+1}-2}{D}\,\|\mathbf{x}\|_2^{2p}\,\|\mathbf{y}\|_2^{2p}. \quad (12)$$

Time Complexity. To compute the sketch according to Equation (7), each complex `CountSketch` can be computed in $O(\mathrm{nnz}\,(\mathbf{x}))$ time. The convolution of the p sketches is implemented using FFT in $O(pD\log D)$ time. Therefore, the overall time complexity of Complex-to-Real (CtR) `TensorSketch` applied to $\mathbf{x}^{\otimes p}$ is $O\big(p(\mathrm{nnz}\,(\mathbf{x})+D\log D)\big)$. □

Remark 3. *Theorem 2, exhibits an improved exponential dependence on p in the variance bound of the estimate, decreasing from the $3^p/D$ bound for real `TensorSketch` [Pham and Pagh, 2013] to $2^p/D$, while*

retaining the same input-sparsity sketching time. A tabular comparison among the baselines on the variance bound and running time is presented in Table 1.

Bounds derived in the Lemma 4 play a key role in the proof of Theorem 2. The sketch described in this lemma is a complex-valued variant of the classical AMS sketch Alon et al. [1999] for polynomial kernel approximation Indyk and McGregor [2008], Braverman et al. [2010]. In particular, the variance analysis of CtR `TensorSketch` reduces to analyzing products of independently randomized linear sketches. Each bucket of CtR `TensorSketch` behaves like a product of p independent AMS-type sketches of the form $Z_{s_j}(\mathbf{x}) = \sum_{i=1}^{d} x_i s_j(i)$. When expanding the moments $\mathbb{E}[|Z|^2]$ and $\mathbb{E}[Z^2]$ in the proof of Theorem 2, the resulting expressions factor across the p independent random variable drawn uniformly from fourth roots of unity. Lemma 4 provides an exact evaluation of moments under fourth roots-of-unity random functions, allowing the full degree-p moment to be written as the p-th power of a single-sketch expression.

Lemma 4. *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 random functions, each taking values uniformly from the four fourth roots of unity (as in Lemma 1). Define*

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

where

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

Then,

$$\mathbb{E}\,[Z] = \langle \mathbf{x}, \mathbf{y} \rangle^p,$$
$$\mathrm{Var}\,(Z) \le 2^p\ \|\mathbf{x}\|_2^{2p}\ \|\mathbf{y}\|_2^{2p}.$$

Proof. Let

$$Z = \prod_{j=1}^{p} Z_{s_j}(\mathbf{x})\,\overline{Z_{s_j}(\mathbf{y})}, \qquad Z_{s_j}(\mathbf{x}) = \sum_{i=1}^{d} x_i s_j(i).$$

Expectation. For a fixed j,

$$\mathbb{E}\left[Z_{s_j}(\mathbf{x})\overline{Z_{s_j}(\mathbf{y})}\right] = \sum_{i,k} x_i y_k\, \mathbb{E}\left[s_j(i)\overline{s_j(k)}\right].$$

By Lemma 1, $\mathbb{E}\left[s_j(i)\overline{s_j(k)}\right] = 0$ for $i \neq k$ and equals 1 for $i = k$. Hence,

$$\mathbb{E}\left[Z_{s_j}(\mathbf{x})\overline{Z_{s_j}(\mathbf{y})}\right] = \langle \mathbf{x}, \mathbf{y} \rangle.$$

Independence across j then gives

$$\mathbb{E}\,[Z] = \prod_{j=1}^{p} \langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle^p.$$

Second moments. Since the sign functions are independent across j,

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

Expanding one factor and using the fourth–moment structure of fourth roots-of-unity random functions (Lemma 1), only index configurations that form pairs survive. The detailed calculation yields moments as follows

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

and

$$\mathbb{E}\left[\left(Z_{s_j}(\mathbf{x})\overline{Z_{s_j}(\mathbf{y})}\right)^2\right] = 2\langle \mathbf{x}, \mathbf{y} \rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2.$$

Therefore, taking p-times product of the terms give moment bounds for Z

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

Variance bound. Using Equation 4, and applying the Cauchy-Schwarz inequality to upper bound $\langle \mathbf{x}, \mathbf{y} \rangle^2 \le \|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2$, together with the fact that $\sum_i x_i^2 y_i^2 \ge 0$, we obtain

$$\mathrm{Var}\,(Z) \le 2^p\ \|\mathbf{x}\|_2^{2p}\ \|\mathbf{y}\|_2^{2p}.$$

□

5 EXPERIMENTS

Experimental setup. We compare our proposed CtR `TensorSketch` against standard `TensorSketch` [Pham and Pagh, 2013]. Along with this we included JL-type polynomial sketches (Gaussian and Rademacher [Kar and Karnick, 2012]) together with their CtR variants [Wacker et al., 2023] for comparison. All methods are implemented by us directly from the original algorithmic descriptions given in the previous sections.

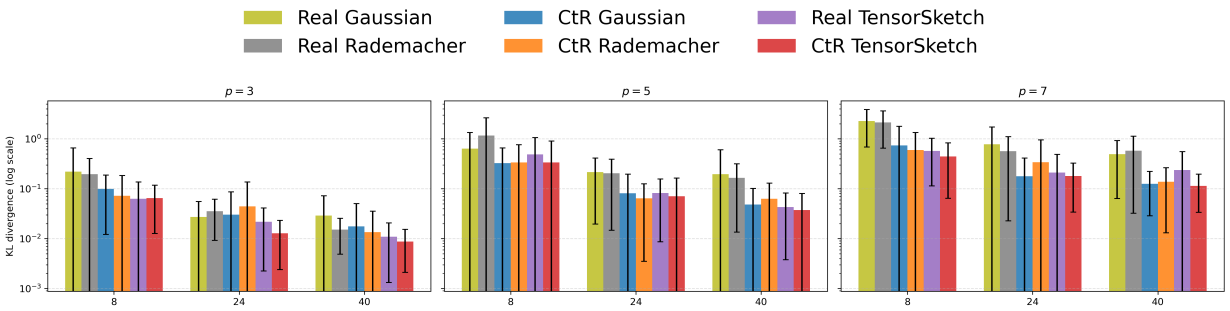

Figure 1: **KL divergence on the COD–RNA dataset.** We report KL divergence between the exact degree-p polynomial kernel and the kernel reconstructed from sketch features for $p \in \{3, 5, 7\}$. Methods include Real and CtR (*complex-to-real*) Gaussian and Rademacher JL sketches, as well as Real and CtR TensorSketch. Results are averaged over 20 independent trials. Sketch dimension is varied as $D \in \{d, 3d, 5d\}$.

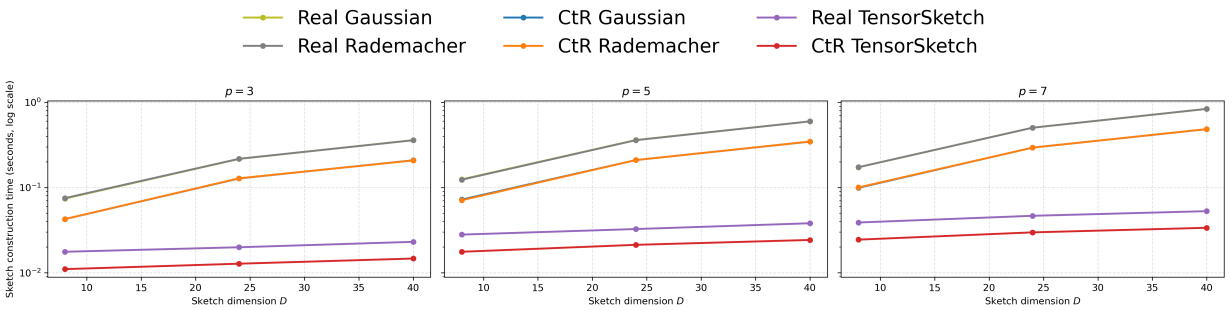

Figure 2: **Wall-clock sketch construction time on the COD–RNA dataset.** We compare Real and CtR (*complex-to-real*) Gaussian and Rademacher JL sketches, with Real and CtR TensorSketch for $p \in \{3, 7, 10\}$. The sketch dimension is varied as $D \in \{d, 3d, 5d\}$. Each point reports the average sketch 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.

All experiments were run on Ubuntu 22.04.4 using an Intel® Core™ i9-14900K processor (24 cores, 32 threads) with 32 GB RAM. Reported runtimes are wall-clock times for sketch construction only.

Baselines. The sketching methods compared in this section are listed below, along with brief descriptions.

- **Real Gaussian/Rademacher Sketch [Kar and Karnick, 2012]:** Dense JL-type polynomial sketches using i.i.d. Gaussian or Rademacher projection matrices.
- **Real `TensorSketch` [Pham and Pagh, 2013, 2025]:** Standard CountSketch-based `TensorSketch` for polynomial kernels with real hash and sign functions.
- **CtR Gaussian/Rademacher Sketch [Wacker et al., 2023]:** *complex-to-real* JL-type sketches using complex Gaussian or complex Rademacher projections followed by CtR conversion, as proposed by [Wacker et al., 2023].
- **CtR `TensorSketch` (Our Proposal):** Our CtR `TensorSketch` obtained by replacing real random signs with random function drawn uniformly from the four fourth roots of unity and applying *complex-to-real* conversion within the corresponding constructions.

Datasets. We use both synthetic and real-world datasets.

Synthetic data. For the synthetic experiments, we generate $n = 3000$ random vectors in $\mathbb{R}^d$ with $d = 2$. Each coordinate is drawn independently from a standard Gaussian distribution, and all vectors are ℓ_2–normalized. We consider polynomial kernels of degrees $p \in \{10, 15, 20\}$ and vary the sketch dimension as $D \in \{100, 300, 500\}$.

Real–world data. We evaluate all methods on two real-world datasets: MAGIC Gamma Telescope [Bock, 2004] ($d = 10$ real-valued features) and COD-RNA [Uzilov et al., 2006] ($d = 8$ numerical attributes). All inputs are treated as real-valued vectors and ℓ_2-normalized prior to sketching. For each dataset, we subsample up to $n = 3000$ points (or use the full dataset if smaller). Each configuration (dataset, degree p, sketch dimension D) is averaged over 20 independent random trials.

Comparison Metrics. We evaluate sketch quality and efficiency using two metrics.

- *KL divergence.* To assess how well a sketch preserves the structure of the degree-p polynomial kernel matrix, we compute the KL divergence between the exact kernel K and its sketch-based approximation $\widehat{K}$. Since polynomial kernels are nonnegative, we normalize both matrices into discrete distributions

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

  and report

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

  where $\varepsilon > 0$ ensures numerical stability. Lower values indicate better kernel preservation.

- *Wall–clock time.* We measure efficiency by the wall–clock time required to construct sketch features for a given method and sketch dimension D, averaged over multiple independent trials.

Results. Figures 1 and 2 compare approximation error (KL divergence) and sketch construction time on the COD–RNA dataset across polynomial degrees and sketch dimensions. We evaluate the proposed CtR `TensorSketch` against the standard `TensorSketch`, along with real and CtR Gaussian and Rademacher JL-type sketches.

CtR `TensorSketch` consistently achieves the lowest KL divergence across all degrees and sketch dimensions, indicating more accurate kernel approximation than real `TensorSketch` and dense JL-type methods. This behavior matches our theoretical guarantee of improved variance established earlier.

In terms of runtime, CtR `TensorSketch` retains the input-sparsity cost $O(p(\mathrm{nnz}\,(\mathbf{x}) + D \log D))$, comparable to standard `TensorSketch`. In contrast, JL-type methods incur higher dense projection cost $O(pDd)$.

Additional experimental results are provided in Section C of the Appendix.

6 CONCLUSION

We introduced a *complex-to-real (CtR)* variant of `TensorSketch` that provides a compression algorithm for high-dimensional polynomial-kernel and tensor datasets. To the best of our knowledge, CtR constructions were only known for dense JL-type `TensorSketch` [Pham and Pagh, 2013] due to Wacker et al. [2023]. Our results demonstrate that Complex-to-Real (CtR) variants achieve improved variance bounds compared to their real counterparts [Pham and Pagh, 2013], matching those obtained in Wacker et al. [2023]. Moreover, these variants preserve the input-sparsity running time of the original method, making it a preferred choice over Wacker et al. [2023], for high-dimensional sparse data.

Two directions remain open for further investigation. First, it is unclear whether employing higher-order roots of unity can yield stronger higher-moment concentration, leading to tighter (ϵ, δ) -approximation guarantees for the sketch. Second, it remains to be explored whether the advantages of Complex-to-Real (CtR) constructions can be extended to other kernel families and randomized feature maps beyond polynomial kernels. Together, these questions point to a broader potential for complex-valued sketch design.

Acknowledgements

This research was partially supported by the ANRF under Project No. ANRF/ECRG/2024/001063/ENS. We gratefully acknowledge ANRF for this support.

References

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1997.1545. URL `https://www.sciencedirect.com/science/article/pii/S0022000097915452`.

Hélène Aschard. A perspective on interaction effects in genetic association studies. *Genetic Epidemiology*, 40 (8):678–688, December 2016. ISSN 0741-0395. doi: 10.1002/gepi.21989. Epub 2016 Jul 7.

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. *Theor. Comput. Sci.*, 312(1):3–15, 2004. doi: 10.1016/S0304-3975(03)00400-6. URL `https://doi.org/10.1016/S0304-3975(03)00400-6`.

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 Jian Su, Kevin Duh, and Xavier

Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas, November 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)*, pages 317–326, Los Alamitos, CA, 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.41. URL https://doi.org/10.1109/CVPR.2016.41.

Yoav Goldberg and Michael Elhadad. splitSVM: Fast, space-efficient, non-heuristic, polynomial kernel computation for NLP applications. In Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, editors, *Proceedings of ACL-08: HLT, Short Papers*, pages 237–240, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL https://aclanthology.org/P08-2060/.

Uffe Haagerup and Magdalena Musat. On the best constants in noncommutative khintchine-type inequalities. *Journal of Functional Analysis*, 250(2):588–624, 2007.

Piotr Indyk and Andrew McGregor. Declaring independence via the sketching of sketches. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '08, page 737–745, USA, 2008. Society for Industrial and Applied Mathematics.

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, pages 2479–2487, Cambridge, MA, 2018. PMLR. URL https://proceedings.mlr.press/v80/kang18b.html.

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

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

Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random fourier features. In *International conference on machine learning*, pages 3905–3914. PMLR, 2019.

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.

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)*, pages 239–247, New York, NY, 2013. ACM. doi: 10.1145/2487575.2487591. URL https://doi.org/10.1145/2487575.2487591.

Ninh Pham and Rasmus Pagh. Tensor sketch: Fast and scalable polynomial kernel approximation, 2025. URL https://arxiv.org/abs/2505.08146.

Rameshwar Pratap and Raghav Kulkarni. Variance reduction in frequency estimators via control variates method. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 161, pages 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)*, pages 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)*, pages 995–1000, Los Alamitos, CA, 2010. IEEE Computer Society. doi: 10.1109/ICDM.2010.127. URL https://doi.org/10.1109/ICDM.2010.127.

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

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.

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, 02 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. *Mach. Learn.*, 111(7):2631–2662, 2022. doi: 10.1007/S10994-022-06166-Z. URL https://doi.org/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. *Inf. Process. Lett.*, 187:106517, 2025. doi: 10.1016/J.IPL.2024.106517. URL https://doi.org/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, pages 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. *J. Mach. Learn. Res.*, 25:235:1–235:75, 2024. URL https://jmlr.org/papers/v25/22-0118.html.

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.

Improving TensorSketch Using Complex Random Variables (Supplementary Material)

Amit Sharma[*1] **Mohammad Azhar Khan**[*1] **Rameshwar Pratap**[1] **Keegan Kang**[2]

[1]Department of Computer Science and Engineering, IIT Hyderabad, India
[2]Department of Mathematics and Statistics, Bucknell University, USA

A PROOFS OF THEOREMS IN SECTION 4

Lemma 1. *Let $\omega = e^{2\pi i/4} = i$, and let $s : [d] \to \{1, \omega, \omega^2, \omega^3\}$ be a random function such that the values $\{s(i)\}_{i\in[d]}$ are drawn independently and uniformly from the four fourth roots of unity. Then, for every $i \in [d]$, the following identities hold*

$$\begin{aligned}\mathbb{E}\left[s(i)\right] &= 0,\\ \mathbb{E}[|s(i)|^2] &= 1,\\ \mathbb{E}\left[s(i)^2\right] &= 0.\end{aligned}$$

Furthermore, for any pair of distinct indices $i \neq j$, independence implies

$$\mathbb{E}[s(i)\,\overline{s(j)}] = \mathbb{E}[s(i)]\,\mathbb{E}\left[\overline{s(j)}\right] = 0.$$

Proof. Since $s(i)$ is uniformly distributed over $\{1, \omega, \omega^2, \omega^3\} = \{1, i, -1, -i\}$, we have

$$\mathbb{E}\left[s(i)\right] = \frac{1 + i - 1 - i}{4} = 0.$$

All four values have unit magnitude, and therefore

$$\mathbb{E}[|s(i)|^2] = \frac{|1|^2 + |i|^2 + |-1|^2 + |-i|^2}{4} = 1.$$

For the second complex moment,

$$\mathbb{E}\left[s(i)^2\right] = \frac{1^2 + i^2 + (-1)^2 + (-i)^2}{4} = \frac{1 - 1 + 1 - 1}{4} = 0.$$

Finally, for any $i \neq j$, independence of $s(i)$ and $s(j)$ implies

$$\mathbb{E}[s(i)\,\overline{s(j)}] = \mathbb{E}[s(i)]\,\mathbb{E}\left[\overline{s(j)}\right] = 0.$$

□

A.1 UNBIASEDNESS AND VARIANCE ANALYSIS OF CTR `TENSORSKETCH`

We establish the theoretical guarantees of CtR `TensorSketch`. Specifically, we show that the estimator is unbiased for the inner product estimation and subsequently provide a variance analysis.

Theorem 2. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p} \in \mathbb{R}^{d^p}$. Let Φ_{CtR} denote a* Complex-to-Real `TensorSketch` *as stated in Definition 8. Let $\widehat{k}_{\mathrm{CtR}}(\mathbf{x}, \mathbf{y}) := \Phi_{\mathrm{CtR}}(\mathbf{x}^{\otimes p})^\top \Phi_{\mathrm{CtR}}(\mathbf{y}^{\otimes p})$, then $\widehat{k}_{\mathrm{CtR}}(\mathbf{x}, \mathbf{y})$ satisfies*

$$\begin{aligned}
\mathbb{E}\left[\widehat{k}_{\mathrm{CtR}}(\mathbf{x}, \mathbf{y})\right] &= \langle \mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle^p, \\
\mathrm{Var}\left(\widehat{k}_{\mathrm{CtR}}(\mathbf{x}, \mathbf{y})\right) &\leq \frac{1}{D}\left(\left(\langle \mathbf{x}, \mathbf{y} \rangle^2 + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 - \sum_{i=1}^d x_i^2 y_i^2\right)^p - \langle \mathbf{x}, \mathbf{y} \rangle^{2p}\right), \\
&\leq \frac{2^p - 1}{D} \|\mathbf{x}\|_2^{2p} \|\mathbf{y}\|_2^{2p}.
\end{aligned}$$

Moreover, the sketches $\Phi_{\mathrm{CtR}}(\mathbf{x}^{\otimes p})$ and $\Phi_{\mathrm{CtR}}(\mathbf{y}^{\otimes p})$ 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. According to Definition 5, we have $\widehat{k}_{\mathrm{CtR}}(\mathbf{x}, \mathbf{y}) := \mathrm{Re}\left\{\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y})\right\}$, and therefore we first analyze $\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y})$. Recall that $\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y}) := \Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p})\overline{\Phi_{\mathrm{C}}(\mathbf{y}^{\otimes p})}^\top$, where $\Phi_{\mathrm{C}}(\mathbf{x}^{\otimes p}) := \mathbf{C}\mathbf{x}^{\otimes p}$ and $\Phi_{\mathrm{C}}(\mathbf{y}^{\otimes p}) := \mathbf{C}\mathbf{y}^{\otimes p}$ according to Definition 8. In particular, $\mathbf{C}\mathbf{x}^{\otimes p}, \mathbf{C}\mathbf{y}^{\otimes p} \in \mathbb{C}^{D/2}$ denote the complex `CountSketch` of the vectors $\mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p} \in \mathbb{R}^{d^p}$, constructed using the derived functions $H : [d]^p \mapsto [D/2]$ and $S : [d]^p \to \{1, \omega, \omega^2, \omega^3\}$ defined as

$$\begin{aligned}
H(i_1, \ldots, i_p) &= \left(\sum_{j=1}^p h_j(i_j)\right) \bmod D/2, \\
S(i_1, \ldots, i_p) &= \prod_{j=1}^p s_j(i_j).
\end{aligned}$$

In the proof, we denote $X := \mathbf{x}^{\otimes p}$ and $Y := \mathbf{y}^{\otimes p}$. Let $u, v \in [d]^p$ be the multi-indices corresponding to the entries of $X, Y \in \mathbb{R}^{d^p}$. For a multi-index $u = (j_1, \ldots, j_p)$, the entry of X is defined as $X_u = \prod_{k=1}^p x_{j_k}$, where the associated linearized index is $u = 1 + \sum_{k=1}^p (j_k - 1) \prod_{k'=1}^{p-k} d$. The entries of Y are defined analogously.

We begin by analyzing $\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y})$ as defined below,

$$\begin{aligned}
\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y}) &:= \Phi_{\mathrm{C}}(X)\overline{\Phi_{\mathrm{C}}(Y)}^\top = \langle \mathbf{C}X, \overline{\mathbf{C}Y} \rangle, \\
&= \sum_{u,v \in [d]^p} X_u Y_v S(u) \overline{S(v)} \mathbf{1}[H(u) = H(v)], \\
&= \langle X, Y \rangle + \sum_{u \neq v} X_u Y_v S(u) \overline{S(v)} \mathbf{1}[H(u) = H(v)].
\end{aligned}$$

Using Lemma 1, we have $\mathbb{E}\left[S(u)\overline{S(v)}\right] = 0$ for all $u \neq v$. Hence,

$$\mathbb{E}\left[\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y})\right] = \langle X, Y \rangle = \langle \mathbf{x}^{\otimes p}, \mathbf{y}^{\otimes p} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle^p.$$

According to Definition 5, we know $\widehat{k}_{\mathrm{CtR}}(\mathbf{x}, \mathbf{y}) := \mathrm{Re}\left\{\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y})\right\}$, we have

$$\mathbb{E}\left[\widehat{k}_{\mathrm{CtR}}(\mathbf{x}, \mathbf{y})\right] = \mathbb{E}\left[\mathrm{Re}\left\{\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y})\right\}\right] = \langle \mathbf{x}, \mathbf{y} \rangle^p. \tag{13}$$

Using Equation 5, the variance of $\widehat{k}_{\mathrm{CtR}}(\mathbf{x}, \mathbf{y})$ is

$$\mathrm{Var}\left(\widehat{k}_{\mathrm{CtR}}(\mathbf{x}, \mathbf{y})\right) = \frac{1}{2}\mathrm{Re}\left\{\mathbb{E}\left[\left|\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y})\right|^2\right] + \mathbb{E}\left[\left(\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y})\right)^2\right] - 2\left(\mathbb{E}\left[\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y})\right]\right)^2\right\}. \tag{14}$$

For the variance analysis, we need to compute $\mathbb{E}\left[\left|\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y})\right|^2\right]$ and $\mathbb{E}\left[\left(\widehat{k}_{\mathrm{C}}(\mathbf{x}, \mathbf{y})\right)^2\right]$.

Therefore by expanding $|\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})|^2$ we get:

$$|\widehat{k}_{\mathrm{C}}(\mathbf{x},\mathbf{y})|^2 := |\langle \mathrm{C}\mathbf{x}^{(p)}, \overline{\mathrm{C}\mathbf{y}^{(p)}}\rangle| = \langle \mathrm{C}\mathbf{x}^{(p)}, \overline{\mathrm{C}\mathbf{y}^{(p)}}\rangle \langle \overline{\mathrm{C}\mathbf{x}^{(p)}}, \mathrm{C}\mathbf{y}^{(p)}\rangle, \tag{15}$$

$$= \left(\langle X, Y\rangle + \sum_{u\neq v} X_u Y_v S(u)\overline{S(v)}\mathbf{1}\left[H(u)=H(v)\right]\right)\left(\langle X, Y\rangle + \sum_{u\neq v} X_u Y_v \overline{S(u)}S(v)\mathbf{1}\left[H(u)=H(v)\right]\right), \tag{16}$$

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

$$+ \left|\left(\sum_{u\neq v} X_u Y_v S(u)\overline{S(v)}\mathbf{1}\left[H(u)=H(v)\right]\right)\right|^2. \tag{17}$$

Therefore, by applying expectation

$$\mathbb{E}\left[\left|\left(\sum_{u\neq v} X_u Y_v S(u)\overline{S(v)}\mathbf{1}\left[H(u)=H(v)\right]\right)\right|^2\right]$$

$$= \mathbb{E}\left[\sum_{\substack{u_1\neq v_1\\ u_2\neq v_2}} X_{u_1} Y_{v_1} X_{u_2} Y_{v_2} S(u_1)\overline{S(v_1)S(u_2)}S(v_2)\mathbf{1}\left[H(u_1)=H(v_1)\right]\mathbf{1}\left[H(u_2)=H(v_2)\right]\right], \tag{18}$$

$$= \sum_{\substack{u_1\neq v_1\\ u_2\neq v_2}} \mathbb{E}\Big[X_{u_1} Y_{v_1} X_{u_2} Y_{v_2} S(u_1)\overline{S(v_1)S(u_2)}S(v_2)\Big]\cdot \mathbb{E}[\mathbf{1}\left[H(u_1)=H(v_1)\right]\mathbf{1}\left[H(u_2)=H(v_2)\right]], \tag{19}$$

$$\leq \frac{2}{D}\sum_{\substack{u_1\neq v_1\\ u_2\neq v_2}} \mathbb{E}\Big[X_{u_1} Y_{v_1} X_{u_2} Y_{v_2} S(u_1)\overline{S(v_1)S(u_2)}S(v_2)\Big], \tag{20}$$

$$\leq \frac{2}{D}\sum_{u_1\neq v_1} \mathbb{E}\Big[|X_{u_1}|\,|Y_{v_1}|\,|X_{u_2}|\,|Y_{v_2}|\,S(u_1)\overline{S(v_1)S(u_2)}S(v_2)\Big], \tag{21}$$

$$= \frac{2}{D}\mathbb{E}\left[\left|\left(\sum_{u\neq v\in[d]^p} |X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\right|^2\right]. \tag{22}$$

We now bound the above expression using the second-moment bound from Lemma 4, given in Equation (66). For completeness, we first restate the bound,

$$\mathbb{E}\left[\left|\left(\sum_{u,v\in[d]^p} |X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\right|^2\right] = \left(\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^p. \tag{23}$$

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 (23) as follows,

$$\left|\left(\sum_{u,v\in[d]^p}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\right|^2 = \left|\sum_{u\in[d]^p}|X_u|\,|Y_v| + \sum_{\substack{u,v\in[d]^p\\u\neq v}}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right|^2, \tag{24}$$

$$= \left(\sum_{u\in[d]^p}|X_u|\,|Y_u| \;+\; \sum_{\substack{u,v\in[d]^p\\u\neq v}}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\overline{\left(\sum_{w\in[d]^p}|X_w|\,|Y_w| \;+\; \sum_{\substack{w,z\in[d]^p\\w\neq z}}|X_w|\,|Y_z|\,S(w)\overline{S(z)}\right)}, \tag{25}$$

$$= \left(\sum_{u\in[d]^p}|X_u|\,|Y_u| \;+\; \sum_{\substack{u,v\in[d]^p\\u\neq v}}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\left(\sum_{w\in[d]^p}|X_w|\,|Y_w| \;+\; \sum_{\substack{w,z\in[d]^p\\w\neq z}}|X_w|\,|Y_z|\,\overline{S(w)}S(z)\right), \tag{26}$$

$$= \sum_{u,w\in[d]^p}|X_u|\,|Y_u|\,|X_w|\,|Y_w| + \sum_{u\in[d]^p}\sum_{\substack{w,z\in[d]^p\\w\neq z}}|X_u|\,|Y_u|\,|X_w|\,|Y_z|\,\overline{S(w)}S(z) + \cdots$$

$$\cdots + \sum_{\substack{u,v\in[d]^p\\u\neq v}}\sum_{w\in[d]^p}|X_u|\,|Y_v|\,|X_w|\,|Y_w|\,S(u)\overline{S(v)} + \sum_{\substack{u,v\in[d]^p\\u\neq v}}\sum_{\substack{w,z\in[d]^p\\w\neq z}}|X_u|\,|Y_v|\,|X_w|\,|Y_z|\,S(u)\overline{S(v)}\,\overline{S(w)}S(z). \tag{27}$$

Using Lemma 1, we have $\mathbb{E}\left[S(u)\,\overline{S(v)}\right] = 0$ for all $u \neq v$. Hence,

$$\sum_{\substack{u,v\in[d]^p\\u\neq v}}\sum_{w\in[d]^p}|X_u|\,|Y_v|\,|X_w|\,|Y_w|\,\mathbb{E}\left[S(u)\overline{S(v)}\right] = 0, \tag{28}$$

Similarly, for all $w \neq z$,

$$\sum_{u\in[d]^p}\sum_{\substack{w,z\in[d]^p\\w\neq z}}|X_u|\,|Y_u|\,|X_w|\,|Y_z|\,\mathbb{E}\left[\overline{S(w)}S(z)\right] = 0, \tag{29}$$

Substitute this in Equation (23), we get

$$\mathbb{E}\left[\left|\sum_{u,v\in[d]^p}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right|^2\right] = \sum_{u,w\in[d]^p}|X_u|\,|Y_u|\,|X_w|\,|Y_w| \;+\cdots$$

$$\cdots + \mathbb{E}\left[\sum_{\substack{u,v\in[d]^p\\u\neq v}}\sum_{\substack{w,z\in[d]^p\\w\neq z}}|X_u|\,|Y_v|\,|X_w|\,|Y_z|\,S(u)\overline{S(v)}\,\overline{S(w)}S(z)\right], \tag{30}$$

$$= \langle X, Y\rangle^2 + \mathbb{E}\left|\left(\sum_{u\neq v}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right)\right|^2. \tag{31}$$

We conclude that,

$$\mathbb{E}\left[\left|\sum_{u\neq v}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right|^2\right] = \mathbb{E}\left[\left|\sum_{u,v\in[d]^p}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right|^2\right] - \langle\mathbf{x},\mathbf{y}\rangle^{2p}. \tag{32}$$

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

$$\mathbb{E}\left[\left|\sum_{u\neq v}|X_u|\,|Y_v|\,S(u)\overline{S(v)}\right|^2\right] = \left(\langle\mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2 - \sum_{i=1}^{d}x_i^2y_i^2\right)^p - \langle\mathbf{x},\mathbf{y}\rangle^{2p}. \tag{33}$$

Put this value in Equation 22,

$$\mathbb{E}\left[\left|\sum_{u\neq v}X_uY_vS(u)\overline{S(v)}\mathbf{1}\left[H(u)=H(v)\right]\right|^2\right] \leq \frac{2}{D}\left(\left(\langle\mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2 - \sum_{i=1}^{d}x_i^2y_i^2\right)^p - \langle\mathbf{x},\mathbf{y}\rangle^{2p}\right). \tag{34}$$

Now we have the second moment as follows,

$$\mathbb{E}\left[|\widehat{k}_{\mathrm{C}}\left(\mathbf{x},\mathbf{y}\right)|^2\right] = \langle X,Y\rangle^2 + \mathbb{E}\left[\left|\left(\sum_{u\neq v}X_uY_vS(u)\overline{S(v)}\mathbf{1}\left[H(u)=H(v)\right]\right)\right|^2\right], \tag{35}$$

$$\leq \langle\mathbf{x},\mathbf{y}\rangle^{2p} + \frac{2}{D}\left(\left(\langle\mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2 - \sum_{i=1}^{d}x_i^2y_i^2\right)^p - \langle\mathbf{x},\mathbf{y}\rangle^{2p}\right). \tag{36}$$

Next, we analyze $\mathbb{E}\Big[(\widehat{k}_{\mathrm{C}}\left(\mathbf{x},\mathbf{y}\right))^2\Big]$. Expanding $(\widehat{k}_{\mathrm{C}}\left(\mathbf{x},\mathbf{y}\right))^2$ gives:

$$(\widehat{k}_{\mathrm{C}}\left(\mathbf{x},\mathbf{y}\right))^2 := \left(\langle\mathrm{C}\mathbf{x}^{(p)},\overline{\mathrm{C}\mathbf{y}^{(p)}}\rangle\right)^2 = \langle\mathrm{C}\mathbf{x}^{(p)},\overline{\mathrm{C}\mathbf{y}^{(p)}}\rangle\langle\mathrm{C}\mathbf{x}^{(p)},\overline{\mathrm{C}\mathbf{y}^{(p)}}\rangle, \tag{37}$$

$$= \left(\langle X,Y\rangle + \sum_{u\neq v}X_uY_vS(u)\overline{S(v)}\mathbf{1}\left[H(u)=H(v)\right]\right)\left(\langle X,Y\rangle + \sum_{u\neq v}X_uY_vS(u)\overline{S(v)}\mathbf{1}\left[H(u)=H(v)\right]\right), \tag{38}$$

$$= \langle X,Y\rangle^2 + 2\langle X,Y\rangle\left(\sum_{u\neq v}X_uY_vS(u)\overline{S(v)}\mathbf{1}\left[H(u)=H(v)\right]\right) + \left(\sum_{u\neq v}X_uY_vS(u)\overline{S(v)}\mathbf{1}\left[H(u)=H(v)\right]\right)^2, \tag{39}$$

Therefore, by applying expectation

$$\mathbb{E}\left[\left(\sum_{u\neq v}X_uY_vS(u)\overline{S(v)}\mathbf{1}\left[H(u)=H(v)\right]\right)^2\right] \tag{40}$$

$$= \mathbb{E}\left[\sum_{\substack{u_1\neq v_1\\u_2\neq v_2}}X_{u_1}Y_{v_1}X_{u_2}Y_{v_2}S(u_1)\overline{S(v_1)}S(u_2)\overline{S(v_2)}\mathbf{1}\left[H(u_1)=H(v_1)\right]\mathbf{1}\left[H(u_2)=H(v_2)\right]\right], \tag{41}$$

$$= \sum_{\substack{u_1\neq v_1\\u_2\neq v_2}}\mathbb{E}\Big[X_{u_1}Y_{v_1}X_{u_2}Y_{v_2}S(u_1)\overline{S(v_1)}S(u_2)\overline{S(v_2)}\Big]\cdot\mathbb{E}[\mathbf{1}\left[H(u_1)=H(v_1)\right]\mathbf{1}\left[H(u_2)=H(v_2)\right]], \tag{42}$$

$$\leq \frac{2}{D}\mathbb{E}\left[\left(\sum_{u\neq v\in[d]^p}X_u\,Y_v\,S(u)\overline{S(v)}\right)^2\right]. \tag{43}$$

By similar analysis and using second-moment bound of Lemma 4 given in Equation (75), we get

$$\mathbb{E}\left[\left(\sum_{u\neq v}X_uY_vS(u)\overline{S(v)}\mathbf{1}\left[H(u)=H(v)\right]\right)^2\right] \leq \frac{2}{D}\left(\left(2\langle\mathbf{x},\mathbf{y}\rangle^2 - \sum_{i=1}^{d}x_i^2y_i^2\right)^p - \langle\mathbf{x},\mathbf{y}\rangle^{2p}\right). \tag{44}$$

Therefore, we have,

$$\mathbb{E}\left[\left(\widehat{k}_{\mathrm{C}}\left(\mathbf{x},\mathbf{y}\right)\right)^2\right] = \langle X,Y\rangle^2 + \mathbb{E}\left[\left(\sum_{u\neq v} X_u Y_v S(u)\overline{S(v)}\mathbf{1}\left[H(u)=H(v)\right]\right)^2\right], \tag{45}$$

$$\leq \langle \mathbf{x},\mathbf{y}\rangle^{2p} + \frac{2}{D}\left(\left(2\langle \mathbf{x},\mathbf{y}\rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^p - \langle \mathbf{x},\mathbf{y}\rangle^{2p}\right). \tag{46}$$

Therefore for final variance we put Equation (22) and Equation (46) in Equation (14),we get

$$\mathrm{Var}\left(\widehat{k}_{\mathrm{C}}\left(\mathbf{x},\mathbf{y}\right)\right) \leq \frac{1}{2}\Bigg[\langle \mathbf{x},\mathbf{y}\rangle^{2p} + \frac{2}{D}\Big(\big(\langle \mathbf{x},\mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2\big)^p - \langle \mathbf{x},\mathbf{y}\rangle^{2p}\Big)$$
$$+ \langle \mathbf{x},\mathbf{y}\rangle^{2p} + \frac{2}{D}\Big(\big(2\langle \mathbf{x},\mathbf{y}\rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2\big)^p - \langle \mathbf{x},\mathbf{y}\rangle^{2p}\Big) - 2\langle \mathbf{x},\mathbf{y}\rangle^{2p}\Bigg], \tag{47}$$

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

$$\leq \frac{1}{2}\left[\frac{2}{D}\left(\left(2\|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2\right)^p - \|\mathbf{x}\|_2^{2p}\,\|\mathbf{y}\|_2^{2p}\right) + \frac{2}{D}\left(\left(2\|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2\right)^p - \|\mathbf{x}\|_2^{2p}\,\|\mathbf{y}\|_2^{2p}\right)\right], \tag{48}$$

$$\leq \frac{1}{2}\left[\frac{2^{p+1}-2}{D}\|\mathbf{x}\|_2^{2p}\,\|\mathbf{y}\|_2^{2p} + \frac{2^{p+1}-2}{D}\|\mathbf{x}\|_2^{2p}\,\|\mathbf{y}\|_2^{2p}\right], \tag{49}$$

$$\leq \frac{2^{p+1}-2}{D}\|\mathbf{x}\|_2^{2p}\,\|\mathbf{y}\|_2^{2p}. \tag{50}$$

□

We begin by stating a lemma that serves as a key step to bound the second moments of the estimator in the proof of Theorem 2 .

Lemma 4. *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 random functions, each taking values uniformly from the four fourth roots of unity (as in Lemma 1). Define*

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

where

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

Then,

$$\mathbb{E}\left[Z\right] = \langle \mathbf{x},\mathbf{y}\rangle^p,$$
$$\mathrm{Var}\left(Z\right) \leq 2^p\;\|\mathbf{x}\|_2^{2p}\;\|\mathbf{y}\|_2^{2p}.$$

Proof. First, we consider the expectation. For each j, we note that

$$\mathbb{E}\Big[Z_{s_j}(\mathbf{x})\,\overline{Z_{s_j}(\mathbf{y})}\Big] = \mathbb{E}\left[\left(\sum_{i=1}^{d} x_i\, s_j(i)\right)\left(\sum_{k=1}^{d} y_k\,\overline{s_j(k)}\right)\right], \tag{51}$$

$$= \sum_{i=1}^{d}\sum_{k=1}^{d} x_i y_k\,\mathbb{E}[s_j(i)\overline{s_j(k)}], \tag{52}$$

$$= \sum_{i=1}^{d} x_i y_i\,\mathbb{E}[|s_j(i)|^2] + \sum_{i\neq k} x_i y_k\,\mathbb{E}[s_j(i)\overline{s_j(k)}], \tag{53}$$

$$= \langle \mathbf{x}, \mathbf{y}\rangle, \tag{54}$$

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

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

$$\mathbb{E}[Z] = \prod_{j=1}^{p} \mathbb{E}[Z_{s_j}(\mathbf{x})\overline{Z_{s_j}(\mathbf{y})}] = \langle \mathbf{x}, \mathbf{y}\rangle^p. \tag{55}$$

Next, to bound the variance,

$$\mathrm{Var}\,(Z) = \frac{1}{2}\,\mathrm{Re}\big\{\mathbb{E}[|Z|^2] + \mathbb{E}[(Z)^2] - 2|\mathbb{E}[Z]|^2\big\}. \tag{56}$$

Because hash functions are independent across different j, we may write

$$\mathbb{E}[|Z|^2] = \prod_{j=1}^{p} \mathbb{E}\Big[|\left(Z_{s_j}(\mathbf{x})\,\overline{Z_{s_j}(\mathbf{y})}\right)|^2\Big] \text{ and } \mathbb{E}[(Z)^2] = \prod_{j=1}^{p} \mathbb{E}\left[\left(Z_{s_j}(\mathbf{x})\,\overline{Z_{s_j}(\mathbf{y})}\right)^2\right] \tag{57}$$

For each j, expanding the square gives

$$\mathbb{E}\Big[|\left(Z_{s_j}(\mathbf{x})\,\overline{Z_{s_j}(\mathbf{y})}\right)|^2\Big] = \mathbb{E}\left[\left(\sum_{i=1}^{d} x_i\, s_j(i)\right)\left(\sum_{k=1}^{d} y_k\,\overline{s_j(k)}\right)\left(\sum_{i=1}^{d} x_i\,\overline{s_j(i)}\right)\left(\sum_{k=1}^{d} y_k\, s_j(k)\right)\right], \tag{58}$$

$$= \sum_{i=1}^{d}\sum_{i'=1}^{d}\sum_{k=1}^{d}\sum_{k'=1}^{d} x_i\, x_{i'}\, y_k\, y_{k'}\,\mathbb{E}\Big[s_j(i)\overline{s_j(k)s_j(i')}s_j(k')\Big]. \tag{59}$$

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

$$\mathbb{E}[s_j(i)\overline{s_j(k)s_j(i')}s_j(k')] = \begin{cases} 1, & \text{if } i = k = i' = k', \\ 1, & \text{if } i = k \neq i' = k', \\ 1, & \text{if } i = i' \neq k = k', \\ 0, & \text{otherwise.} \end{cases} \tag{60}$$

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

$$\sum_{i=1}^{d} x_i^2 y_i^2. \tag{61}$$

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

$$\sum_{i\neq i'} x_i y_i\, x_{i'} y_{i'} = \left(\sum_{i=1}^{d} x_i y_i\right)^2 - \sum_{i=1}^{d} x_i^2 y_i^2 = \langle \mathbf{x}, \mathbf{y}\rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2. \tag{62}$$

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

$$\sum_{i \neq k} x_i^2 y_k^2 = \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2. \tag{63}$$

Thus, summing these contributions, we have

$$\mathbb{E}\big[\,|\,(Z_{s_j}(\mathbf{x})\, Z_{s_j}(\mathbf{y}))\,|^2\big] = \sum_{i=1}^{d} x_i^2 y_i^2 + \left(\langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - 2\sum_{i=1}^{d} x_i^2 y_i^2\right), \tag{64}$$

$$= \langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2. \tag{65}$$

Substituting this bound into (1) yields

$$\mathbb{E}[|Z|^2] = \left(\langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^p, \tag{66}$$

For each j, expanding the square gives

$$\mathbb{E}\left[\left(Z_{s_j}(\mathbf{x})\,\overline{Z_{s_j}(\mathbf{y})}\right)^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^{d} x_i\, s_j(i)\right)\left(\sum_{k=1}^{d} y_k\,\overline{s_j(k)}\right)\left(\sum_{i=1}^{d} x_i\, s_j(i)\right)\left(\sum_{k=1}^{d} y_k\,\overline{s_j(k)}\right)\right], \tag{67}$$

$$= \sum_{i=1}^{d}\sum_{i'=1}^{d}\sum_{k=1}^{d}\sum_{k'=1}^{d} x_i\, x_{i'}\, y_k\, y_{k'}\, \mathbb{E}\Big[s_j(i)\overline{s_j(k)}s_j(i')\overline{s_j(k')}\Big]. \tag{68}$$

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

$$\mathbb{E}[s_j(i)\overline{s_j(k)}s_j(i')\overline{s_j(k')}] = \begin{cases} 1, & \text{if } i = k = i' = k', \\ 1, & \text{if } i = k \neq i' = k', \\ 1, & \text{if } i = k' \neq i' = k, \\ 0, & \text{otherwise.} \end{cases} \tag{69}$$

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

$$\sum_{i=1}^{d} x_i^2 y_i^2. \tag{70}$$

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

$$\sum_{i \neq i'} x_i y_i\, x_{i'} y_{i'} = \left(\sum_{i=1}^{d} x_i y_i\right)^2 - \sum_{i=1}^{d} x_i^2 y_i^2 = \langle \mathbf{x}, \mathbf{y}\rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2. \tag{71}$$

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

$$\sum_{i \neq i'} x_i y_i\, x_{i'} y_{i'} = \left(\sum_{i=1}^{d} x_i y_i\right)^2 - \sum_{i=1}^{d} x_i^2 y_i^2 = \langle \mathbf{x}, \mathbf{y}\rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2. \tag{72}$$

Thus, summing these contributions, we have

$$\mathbb{E}\Big[\big(Z_{s_j}(\mathbf{x})\, Z_{s_j}(\mathbf{y})\big)^2\Big] = \sum_{i=1}^{d} x_i^2 y_i^2 + \left(2\langle \mathbf{x}, \mathbf{y}\rangle^2 - 2\sum_{i=1}^{d} x_i^2 y_i^2\right), \tag{73}$$

$$= 2\langle \mathbf{x}, \mathbf{y}\rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2. \tag{74}$$

Substituting this bound into (1) yields

$$\mathbb{E}[(Z)^2] = \left(2\langle \mathbf{x}, \mathbf{y}\rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^p, \tag{75}$$

which completes the proof since

$$\begin{aligned}\operatorname{Var}(Z) &= \frac{1}{2}\operatorname{Re}\big\{\mathbb{E}[|Z|^2] + \mathbb{E}[(Z)^2] - 2|\mathbb{E}[Z]|^2\big\}, &(76)\\ &= \frac{1}{2}\left\{\left(\langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^p + \left(2\langle \mathbf{x}, \mathbf{y}\rangle^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^p - 2\langle \mathbf{x}, \mathbf{y}\rangle^{2p}\right\}. &(77)\end{aligned}$$

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

$$\operatorname{Var}(Z) \le 2^p \|\mathbf{x}\|_2^{2p} \|\mathbf{y}\|_2^{2p}. \tag{78}$$

□

B PROOF OF `COUNTSKETCH` ESTIMATOR WITH FOURTH ROOTS OF UNITY

Theorem 5. *Let Φ denote the CountSketch with the hash function taking values in the fourth roots of unity. For any vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the corresponding sketch vector is defined by*

$$\phi(\mathbf{x})_j = \sum_{i=1}^{d} \delta_{ji}\sigma(i)x_i\,, \qquad \phi(\mathbf{y})_j = \sum_{i=1}^{d} \delta_{ji}\sigma(i)y_i, \qquad \forall j \in [D/2]. \tag{79}$$

where,

- $h : [d] \to [D/2]$ *assigning each coordinate independently and uniformly to one of $D/2$ buckets, and*
- $\sigma : [d] \to \{1, \omega, \omega^2, \omega^3\}$ *(Fourth roots of unity),*
- $\delta_{ji} = \begin{cases} 1, & \text{if } h(i) = j, \\ 0, & \text{otherwise.} \end{cases}$

then the estimator $\hat{k}(\mathbf{x}, \mathbf{y}) := \operatorname{Re}\left\{\Phi(\mathbf{x})^\top \overline{\Phi(\mathbf{y})}\right\}$, satisfies

$$\mathbb{E}\Big[\hat{k}(\mathbf{x}, \mathbf{y})\Big] = \langle \mathbf{x}, \mathbf{y}\rangle, \tag{80}$$

$$\operatorname{Var}\big(\hat{k}(\mathbf{x}, \mathbf{y})\big) = \frac{1}{D}\left(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + \langle \mathbf{x}, \mathbf{y}\rangle^2 - 2\sum_{i=1}^{d} x_i^2 y_i^2\right). \tag{81}$$

Proof. Since $\sigma(i)$ is uniformly distributed over $\{1, \omega, \omega^2, \omega^3\} = \{1, i, -1, -i\}$ and satisfies $\mathbb{E}[\sigma(i)] = 0$, $\mathbb{E}\big[\,|\sigma(i)|^2\,\big] = 1$, $\mathbb{E}\big[(\sigma(i))^2\big] = 0$. Consider now the complex inner product of the sketch:

$$\hat{k}_C(\mathbf{x}, \mathbf{y}) := \Phi(\mathbf{x})^\top \overline{\Phi(\mathbf{y})} = \sum_{j=1}^{D/2} \Big(\sum_{i:\, h(i)=j} \sigma(i)x_i\Big)\Big(\sum_{m:\, h(m)=j} \overline{\sigma(m)}\, y_m\Big) \tag{82}$$

$$= \sum_{i=1}^{d} |\sigma(i)|^2 x_i y_i + \sum_{i \ne m} \mathbf{1}\{h(i) = h(m)\}\, \sigma(i)\overline{\sigma(m)}\, x_i y_m. \tag{83}$$

Taking expectation with respect to h and σ, we get

$$\mathbb{E}\Big[\hat{k}_C(\mathbf{x},\mathbf{y})\Big] = \sum_{i=1}^{d} \mathbb{E}\big[|\sigma(i)|^2\big]\, x_i y_i = \sum_{i=1}^{d} x_i y_i = \langle \mathbf{x},\mathbf{y}\rangle. \tag{84}$$

Since the right-hand side of (84) is real, the same unbiasedness holds for the estimator, which takes the real part of the complex sketch:

$$\mathbb{E}\Big[\hat{k}(\mathbf{x},\mathbf{y})\Big] = \mathbb{E}\left[\mathrm{Re}\left\{\hat{k}_C(\mathbf{x},\mathbf{y})\right\}\right] = \mathrm{Re}\left\{\mathbb{E}\left[\hat{k}_C(\mathbf{x},\mathbf{y})\right]\right\} = \langle \mathbf{x},\mathbf{y}\rangle. \tag{85}$$

To find the exact variance of the estimator, we leverage the Complex-to-Real (CtR) framework stated in our paper. Now, we work out with $\mathbb{E}\Big[|\hat{k}_C(\mathbf{x},\mathbf{y})|^2\Big]$.

$$\mathbb{E}\Big[|\hat{k}_C(\mathbf{x},\mathbf{y})|^2\Big] = \mathbb{E}\Big[|\Phi(\mathbf{x})^\top \overline{\Phi(\mathbf{y})}|^2\Big] = \mathbb{E}\left[\left|\sum_{j=1}^{D/2}\left(\sum_{i=1}^{d}\delta_{ji}\sigma(i)x_i\right)\left(\sum_{m=1}^{d}\delta_{jm}\overline{\sigma(m)}\, y_m\right)\right|^2\right], \tag{86}$$

$$= \sum_{j,j'}^{D/2}\sum_{i,m,i',m'}^{d} \mathbb{E}\big[\delta_{ji}\delta_{jm}\delta_{j'i}\delta_{j'm'}\big]\ \mathbb{E}\big[\sigma(i)\overline{\sigma(m)}\,\overline{\sigma(i')}\sigma(m')\big]\ x_i y_m x_{i'} y_{m'}. \tag{87}$$

Hence,

$$\mathbb{E}\big[\sigma(i)\overline{\sigma(m)}\,\overline{\sigma(i')}\sigma(m')\big] \neq 0 \iff \{i,m,i',m'\} \text{ appear in pairs,}$$

1. With surviving index configurations, when $\mathbf{j} = \mathbf{j}'$:
   - $\mathbf{i} = \mathbf{m} = \mathbf{i}' = \mathbf{m}'$: $\mathbb{E}[\delta_{ji}]\mathbb{E}[|\sigma(i)|^4]x_i^2 y_i^2$,
   - $\mathbf{i} = \mathbf{m} \neq \mathbf{i}' = \mathbf{m}'$: $\mathbb{E}[\delta_{ji}\delta_{ji'}]\mathbb{E}[|\sigma(i)|^2|\sigma(i')|^2]x_i y_i x_{i'} y_{i'}$,
   - $\mathbf{i} = \mathbf{i}' \neq \mathbf{m} = \mathbf{m}'$: $\mathbb{E}[\delta_{ji}\delta_{jm}]\mathbb{E}[|\sigma(i)|^2|\sigma(m)|^2]x_i^2 y_m^2$,
2. With surviving index configurations, when $\mathbf{j} \neq \mathbf{j}'$:
   - $\mathbf{i} = \mathbf{m} \neq \mathbf{i}' = \mathbf{m}'$: $\mathbb{E}[\delta_{ji}\delta_{j'i'}]\mathbb{E}[|\sigma(i)|^2|\sigma(i')|^2]x_i y_i x_{i'} y_{i'}$,

Thus,

$$\mathbb{E}\Big[|\hat{k}_C(\mathbf{x},\mathbf{y})|^2\Big] = \sum_{i=1}^{d} x_i^2 y_i^2 + \frac{2}{D}\sum_{i\neq i'}(x_i y_i x_{i'} y_{i'} + x_i^2 y_{i'}^2) + \frac{D-2}{D}\sum_{i\neq i'} x_i y_i x_{i'} y_{i'}, \tag{88}$$

$$= \langle \mathbf{x},\mathbf{y}\rangle^2 + \frac{2}{D}\sum_{i\neq i'} x_i^2 y_{i'}^2. \tag{89}$$

Now, similarly we get

$$\mathbb{E}\Big[(\hat{k}_C(\mathbf{x},\mathbf{y}))^2\Big] = \sum_{i=1}^{d} x_i^2 y_i^2 + \frac{4}{D}\sum_{i\neq i'}(x_i y_i x_{i'} y_{i'}) + \frac{D-2}{D}\sum_{i\neq i'} x_i y_i x_{i'} y_{i'}, \tag{90}$$

$$= \langle \mathbf{x},\mathbf{y}\rangle^2 + \frac{2}{D}\sum_{i\neq i'}(x_i y_i x_{i'} y_{i'}). \tag{91}$$

Now, substitute values of $\mathbb{E}\Big[|\hat{k}_C(\mathbf{x},\mathbf{y})|^2\Big]$ and $\mathbb{E}\Big[(\hat{k}_C(\mathbf{x},\mathbf{y}))^2\Big]$ and simplifying, we get

$$\mathrm{Var}\big(\hat{k}(\mathbf{x},\mathbf{y})\big) = \frac{1}{D}\left(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + \langle \mathbf{x},\mathbf{y}\rangle^2 - 2\sum_{i=1}^{d} x_i^2 y_i^2\right). \tag{92}$$

Hence, from the above variance bound, it follows that `CountSketch` with complex random variables does not provide any variance reduction over its real-valued counterpart. □

C FURTHER EXPERIMENTS

C.1 EXTENDED EVALUATION ON VARIANCE AND TIME IN DIFFERENT SET-UP

In this section, we present additional experimental results that complement the main empirical evaluation reported in Section 5. These experiments extend the comparison between real and *complex-to-real (CtR)* sketching constructions across additional datasets, higher polynomial degrees, and varied embedding dimension. We report both approximation quality (via KL divergence between exact and sketched kernels) and wall-clock sketch construction time. Together, these results provide a more detailed view of the variance and construction time trade-offs of CtR `TensorSketch` relative to its real-valued counterparts and JL-type baselines.

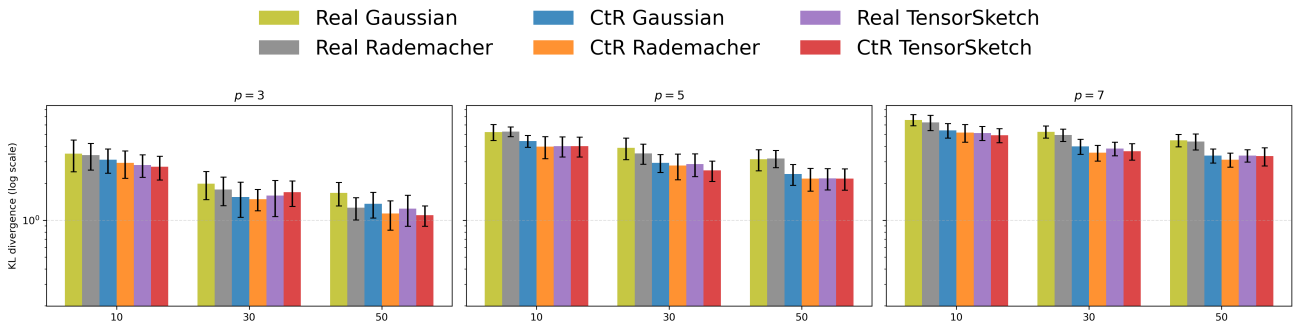

Figure 3: **KL divergence on the MAGIC Gamma Telescope dataset.** We compare Real and *complex-to-real (CtR)* Gaussian and Rademacher JL sketches, together with Real and CtR TensorSketch. Results are shown for polynomial degrees $p \in \{3, 5, 7\}$ and sketch dimensions $D \in \{d, 3d, 5d\}$ with $n = 3000$ standardized and ℓ_2-normalized samples. Bars report the mean KL divergence over 20 independent trials.

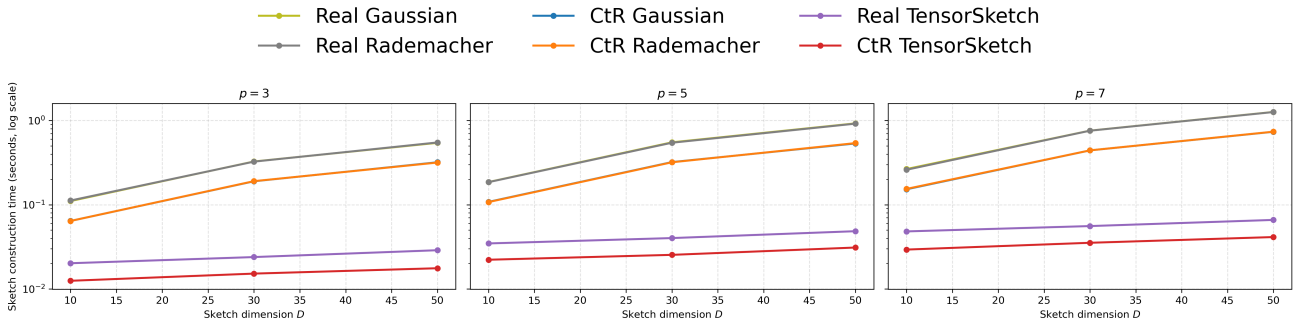

Figure 4: **Wall-clock sketch construction time on the MAGIC dataset.** Methods compared include Real and *complex-to-real (CtR)* Gaussian and Rademacher JL sketches, with Real and CtR TensorSketch. Results are shown for polynomial degrees $p \in \{3, 5, 7\}$ and sketch dimensions $D \in \{d, 3d, 5d\}$ with $n = 3000$ standardized and ℓ_2-normalized samples. Each point reports the mean runtime over 50 independent trials for feature sketch construction only.

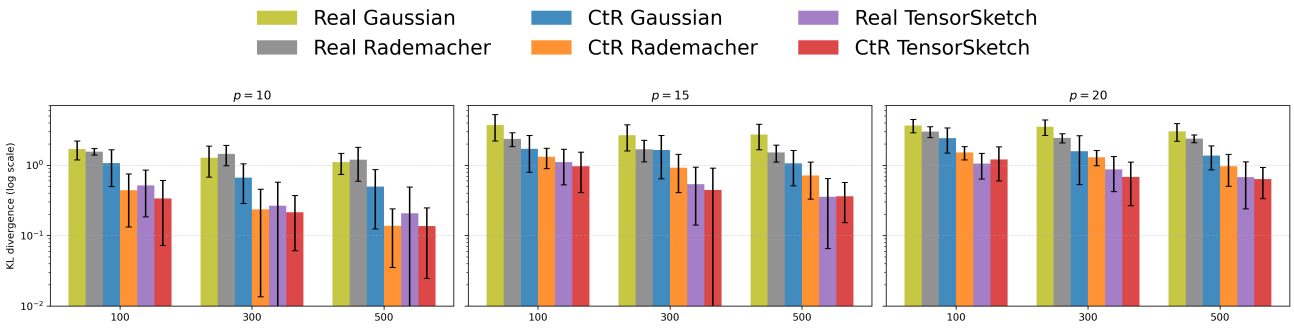

Figure 5: **KL divergence on the synthetic dataset** ($d = 2$)**.** We compare Real and *complex-to-real (CtR)* Gaussian and Rademacher JL sketches, together with Real and CtR TensorSketch, for approximating degree-p polynomial kernels on synthetic Gaussian data ($n = 3000$, dimension $d = 2$, standardized and ℓ_2-normalized). Results are shown for polynomial degrees $p \in \{10, 15, 20\}$ and sketch dimensions $D \in \{100, 300, 500\}$. Bars report the mean KL divergence between the exact kernel and the sketch-based approximation over 20 independent trials.

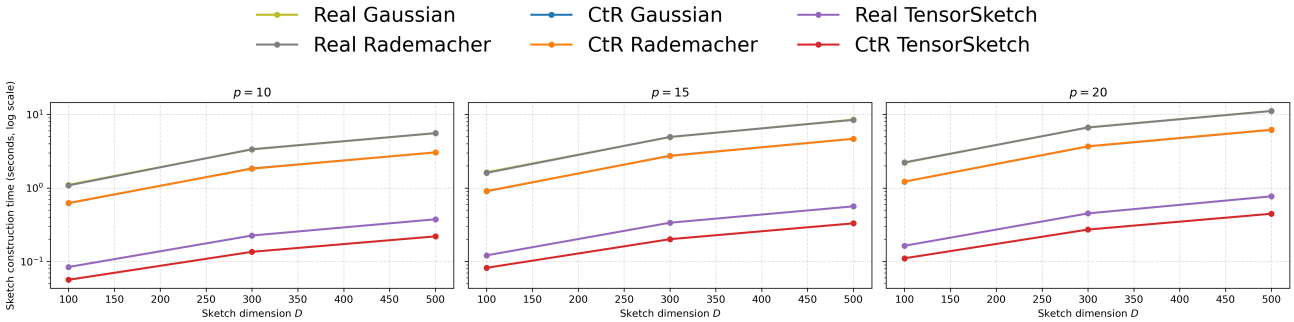

Figure 6: **Wall-clock sketch construction time on a synthetic dataset.** Methods compared include Real and *complex-to-real (CtR)* Gaussian and Rademacher JL sketches, together with Real and CtR TensorSketch. Results are shown for polynomial degrees $p \in \{10, 15, 20\}$ and sketch dimensions $D \in \{100, 300, 500\}$ with $n = 3000$ standardized and ℓ_2-normalized samples in dimension $d = 2$. Each point reports the mean runtime over 20 independent trials for feature sketch construction only (log-scale on the y-axis).

C.2 EVALUATION ON FROBENIUS NORMALIZED RELATIVE ERROR

We also test our proposed method along with other baselines using the Frobenius normalized relative error, which is defined as $\|K - \hat{K}\|_F / \|K\|_F$. In this formula, K represents the exact kernel matrix, while $\hat{K}$ represents the approximated kernel matrix produced by the sketching method. Frobenius error metric is a standard benchmark in kernel approximation and is widely used to measure sketching accuracy Wacker et al. [2024, 2023]. The datasets and experimental setups used here are identical to our previous experiments that evaluated accuracy using the KL divergence metric (Section 5 and Appendix C.1). Testing with this alternative metric ensures that our method performs well irrespective of the choice of error metric.

Degree (p)	**Method**	**Variance**
15	Real Gaussian	4.41×10^{4}
	Real Rademacher	3.20×10^{1}
	CtR Gaussian	6.35×10^{1}
	CtR Rademacher	4.16×10^{3}
	Real TensorSketch	2.84×10^{-5}
	CtR TensorSketch (Ours)	$\mathbf{1.57 \times 10^{-5}}$
20	Real Gaussian	3.80×10^{1}
	Real Rademacher	6.78×10^{3}
	CtR Gaussian	2.76×10^{1}
	CtR Rademacher	1.58×10^{1}
	Real TensorSketch	1.28×10^{-2}
	CtR TensorSketch (Ours)	$\mathbf{5.05 \times 10^{-5}}$
25	Real Gaussian	2.65×10^{0}
	Real Rademacher	9.06×10^{3}
	CtR Gaussian	1.82×10^{1}
	CtR Rademacher	2.58×10^{1}
	Real TensorSketch	3.68×10^{-1}
	CtR TensorSketch (Ours)	$\mathbf{3.58 \times 10^{-4}}$
30	Real Gaussian	7.24×10^{0}
	Real Rademacher	1.00×10^{0}
	CtR Gaussian	5.44×10^{2}
	CtR Rademacher	3.21×10^{3}
	Real TensorSketch	1.27×10^{-3}
	CtR TensorSketch (Ours)	$\mathbf{4.59 \times 10^{-6}}$

Table 2: Variance of the Frobenius normalized relative error ($\|K - \hat{K}\|_F / \|K\|_F$) evaluated on the real-world MAGIC dataset ($n = 1000, d = 10$) with a compressed sketch dimension of $D = 128$. Results are aggregated over 20 independent trials across high-degree polynomial kernels ($p \in \{15, 20, 25, 30\}$). CtR TensorSketch strictly outperforms Real TensorSketch in estimation stability, confirming the variance reduction.

C.3 EVALUATION ON DOWNSTREAM TASKS

To demonstrate practical application beyond kernel matrix approximation, we evaluate our method on downstream binary classification tasks using a linear Support Vector Machine (SVM). The datasets used in these experiments are generated using standard ML libraries (such as scikit-learn) to create two highly overlapping classes with non-linear decision boundaries. We test two separate configurations under extreme compression to a sketch dimension of $D = 64$.

We compare the sketching methods against the Exact Polynomial Kernel, which computes the full kernel matrix using the mathematical formula $K(x, y) = (x^\top y)^p$ without any compression or approximation. We evaluate the techniques using two metrics: test classification accuracy and total execution time. The time metric tracks the combined end-to-end wall-clock time required for both data compression and the subsequent classifier training. We report this total time to fully reflect the complete workload required to obtain the final classification model.

As shown in Table 3 and Table 4, all approximation methods experience a drop in accuracy compared to the exact kernel due

to the tight bottleneck of the compressed dimension. However, our proposed CtR `TensorSketch` consistently achieves the highest accuracy among all baselines while requiring the shortest total execution time.

Method	Accuracy	Time (s)
Exact Poly Kernel	0.8756	16.0724
TensorSketch (Real)	0.4911	0.2390
TensorSketch (CtR) [Ours]	**0.5289**	**0.1479**
JL (CtR Rademacher)	0.5011	3.8279
JL (CtR Gaussian)	0.4978	3.7708

Table 3: Downstream Classification Performance (Linear SVM) for **Setup 1**. Evaluated on a synthetic dataset ($n = 3000$) with dimension $d = 20$. We approximate a polynomial kernel of degree $p = 15$ using a sketch dimension of $D = 64$. Our proposed method achieves the best sketching accuracy in the shortest time.

Method	Accuracy	Time (s)
Exact Poly Kernel	0.8567	8.5344
TensorSketch (Real)	0.4978	0.2763
TensorSketch (CtR) [Ours]	**0.5044**	**0.1733**
JL (CtR Rademacher)	0.4567	3.3204
JL (CtR Gaussian)	0.4889	3.2954

Table 4: Downstream Classification Performance (Linear SVM) for **Setup 2**. Evaluated on a synthetic dataset ($n = 3000$) with dimension $d = 10$. We approximate a polynomial kernel of higher degree $p = 25$ using a sketch dimension of $D = 64$.

C.4 VARIANCE ANALYSIS THROUGH NUMERICAL TABLE

The primary purpose of this section is to provide a clear and direct validation of our experimental findings from Section 5. In our main evaluation, we rely on visual plots to demonstrate the accuracy and computational time of our proposed CtR `TensorSketch`. Graphs especially those that use a logarithmic scale, can visually compress the performance gap between methods. Looking at the exact numbers, we can clearly confirm our previous experiments and highlight the advantage of our method. Specifically, these numbers reveal how our approach successfully reduces the variance from $3^p/D$ in traditional methods to $2^p/D$ in our complex-to-real design.

As shown in Table 5, our proposed CtR `TensorSketch` consistently achieves lower variance compared to all the baselines across all degrees.

Degree (p)	Method	Variance
15	Real Gaussian	0.2634
	Real Rademacher	0.3839
	CtR Gaussian	0.2192
	CtR Rademacher	0.2029
	Real TensorSketch	0.5496
	CtR TensorSketch (Ours)	**0.1739**
20	Real Gaussian	0.3545
	Real Rademacher	0.4673
	CtR Gaussian	0.6996
	CtR Rademacher	0.5246
	Real TensorSketch	0.3932
	CtR TensorSketch (Ours)	**0.3888**
25	Real Gaussian	0.3158
	Real Rademacher	0.2560
	CtR Gaussian	0.4395
	CtR Rademacher	0.2294
	Real TensorSketch	0.3209
	CtR TensorSketch (Ours)	**0.2367**
30	Real Gaussian	0.4785
	Real Rademacher	0.2932
	CtR Gaussian	0.3266
	CtR Rademacher	0.2883
	Real TensorSketch	0.2443
	CtR TensorSketch (Ours)	**0.1939**

Table 5: Variance of KL Divergence for high-degree polynomial kernels. Evaluated on a positive orthant synthetic dataset ($n = 1000, d = 10$) with extreme compression ($D = 64$), aggregated over 20 independent trials. CtR TensorSketch strictly dominates Real TensorSketch in estimation variance across all polynomial degrees, empirically validating the theoretical $2^p/D$ scaling advantage.