Complex-to-Real Sketches for Tensor Products with Applications to the Polynomial Kernel

Jonas Wacker, Ruben Ohana, Maurizio Filippone

Complex-to-Real Sketches for Tensor Products with Applications to the Polynomial Kernel

Jonas Wacker
EURECOM, France

Ruben Ohana
CCM, Flatiron Institute, USA

Maurizio Filippone
EURECOM, France

Abstract

Randomized sketches of a tensor product of p vectors follow a tradeoff between statistical efficiency and computational acceleration. Commonly used approaches avoid computing the high-dimensional tensor product explicitly, resulting in a suboptimal dependence of $\mathcal{O}(3^p)$ in the embedding dimension. We propose a simple Complex-to-Real (CtR) modification of well-known sketches that replaces real random projections by complex ones, incurring a lower $\mathcal{O}(2^p)$ factor in the embedding dimension. The output of our sketches is real-valued, which renders their downstream use straightforward. In particular, we apply our sketches to p-fold self-tensored inputs corresponding to the feature maps of the polynomial kernel. We show that our method achieves state-of-the-art performance in terms of accuracy *and* speed compared to other randomized approximations from the literature.

1 INTRODUCTION

Randomized linear sketching (Woodruff, 2014) is a computationally efficient method for dimensionality reduction, where an input point $\boldsymbol{x} \in \mathbb{R}^d$ is multiplied by a random D-by-d matrix $\boldsymbol{S}$ to yield a low-distortion embedding. When $D \ll d$, the sketched data is more compact, accelerating downstream learning algorithms with statistical guarantees. It is well-known that an optimal choice of $\boldsymbol{S}$ requires an embedding dimension $D = \Theta(\log(1/\delta)\epsilon^{-2})$ to guarantee that $\|\boldsymbol{S}\boldsymbol{x}\|_2$ lies within $(1\pm\epsilon)\,\|\boldsymbol{x}\|_2$ with probability at least $1-\delta$ (Larsen & Nelson, 2017).

Here we consider sketches of tensor products $\otimes_{i=1}^p \boldsymbol{x}_i$ for some arbitrary vectors $\boldsymbol{x}_1 \in \mathbb{R}^{d_1}, \dots, \boldsymbol{x}_p \in \mathbb{R}^{d_p}$. Storing $\otimes_{i=1}^p \boldsymbol{x}_i$ takes $\mathcal{O}(\prod_{i=1}^p d_i)$ memory and becomes infeasible when p or $\{d_i\}_{i=1}^p$ are moderately large, impeding the construction of an explicit sketch. To solve this problem, *implicit* sketching methods have been developed in the past (e.g., Kar & Karnick, 2012; Pham & Pagh, 2013) that compute $\boldsymbol{S}(\otimes_{i=1}^p \boldsymbol{x}_i)$ without ever forming $\otimes_{i=1}^p \boldsymbol{x}_i$.

Sketches for tensor products have been successfully applied to compress deep neural networks for the tasks of fine-grained visual recognition (Gao et al., 2016) and multimodal fusion (Fukui et al., 2016). Furthermore, when considering the special case of self-tensored inputs (we set $\boldsymbol{x} := \boldsymbol{x}_1 = \cdots = \boldsymbol{x}_p$), then $\otimes_{i=1}^p \boldsymbol{x}_i$ corresponds to the feature map of the polynomial kernel. For two inputs $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, the sketch thus yields a randomized approximation $\hat{k}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{S}(\otimes_{i=1}^p \boldsymbol{x}))^\top \boldsymbol{S}(\otimes_{i=1}^p \boldsymbol{y})$ of the polynomial kernel $k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^\top \boldsymbol{y})^p$. This observation connects these sketching methods to random feature maps originally proposed for shift-invariant kernels (Rahimi & Recht, 2007). Polynomial kernels are among the most popular kernels and have proven effective in applications such as natural language processing (Goldberg & Elhadad, 2008), recommender systems (Rendle, 2010), and genomic data analysis (Aschard, 2016). Moreover, more general dot product kernels can be formulated as a positively weighted sum of polynomial kernels through a Taylor expansion (Kar & Karnick, 2012). An extended version of this expansion also exists for the Gaussian kernel (Cotter et al., 2011).

Although it is of high interest to accelerate the aforementioned applications via sketching, commonly used methods proposed in the past require a suboptimal embedding dimension $D = \mathcal{O}(3^p \log(1/\delta)\epsilon^{-2})$ as shown by Avron et al. (2014) and Ahle et al. (2020, Appendix A.2), thus trading statistical efficiency for computational accelerations. Ahle et al. (2020) improve the dependence on p to polynomial by composing well-known base sketches, but require a more expensive meta-algorithm (Song et al., 2021).

In this work, we address this issue from another angle by studying simple complex-valued modifications of existing sketches. These can yield much lower variances as shown in Wacker et al. (2022), but may render a downstream task such as ridge regression more expensive due to linear algebra operations being applied to complex data. More-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

over, Wacker et al. (2022) do not provide guarantees on the preservation of the L2-norm, nor do they provide an intuitive explanation for the improved statistical properties of such sketches. In this sense, our work continues where the previous work falls short. We show that complex sampling distributions have smaller higher-order moments than real-valued analogs while also yielding valid sketches, and we provide an in-depth analysis of resulting theoretical guarantees. We further show that a concatenation of the real and imaginary parts of a complex sketch inherits its statistical advantages and we call the real-valued result a *Complex-to-Real (CtR)* sketch. CtR-sketches are simple to construct and can be used in any downstream task without requiring the model to handle complex data.

More precisely, we make the following main contributions: 1) In Section 3.1, we show that complex sketches preserve the L2-norm of an input vector using only $D = \mathcal{O}(2^p)$ instead of $D = \mathcal{O}(3^p)$ required by their real analogs, while explaining the intuition for this improvement. 2) In Section 3.2, we show that these results readily extend to CtR-sketches resulting in the same guarantees for the approximate matrix product. 3) In Section 3.3, we focus on polynomial kernels and derive the variances of kernel approximations obtained through CtR-sketches, while comparing them against real-valued analogs. 4) In Section 6, we empirically compare a newly developed structured CtR-sketch against the state-of-the-art.

We made the code for this work publicly available.[1]

2 PRELIMINARIES

Notation We denote the tensor product of two vectors $\boldsymbol{a}, \boldsymbol{b}$ as $\boldsymbol{a} \otimes \boldsymbol{b} = \mathrm{vec}(\boldsymbol{a}\boldsymbol{b}^\top)$. For p vectors $\{\boldsymbol{a}_i\}_{i=1}^p$, we use $\otimes_{i=1}^p \boldsymbol{a}_i$. In particular, we write $\boldsymbol{a}^{\otimes p} := \otimes_{i=1}^p \boldsymbol{a}$ when this operation is applied to a vector with itself. For two matrices $\boldsymbol{A}, \boldsymbol{B}$, we denote their element-wise product as $\boldsymbol{A} \odot \boldsymbol{B}$. When they are positive semi-definite (psd), we write $\boldsymbol{A} \preceq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is psd. The Frobenius norm is defined as $\|\boldsymbol{A}\|_F = (\sum_{i,j} A_{i,j}^2)^{1/2}$. For a random variable X, we denote its expected value by $\mathbb{E}[X]$ and its variance by $\mathbb{V}[X]$. Its L^t-norm is $\|X\|_{L^t} = \mathbb{E}[|X|^t]^{1/t}$ for $t \geq 1$.

We define $\mathrm{i} := \sqrt{-1}$. The real-valued standard normal distribution is defined as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and the complex one as $\mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$. The real Rademacher distribution is denoted by $\mathrm{Unif}(\{1, -1\})$, and the complex one by $\mathrm{Unif}(\{1, -1, \mathrm{i}, -\mathrm{i}\})$. A Rademacher vector has its elements drawn i.i.d. from the Rademacher distribution.

Polynomial kernel In this work, we consider polynomial kernels of the form

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\gamma \boldsymbol{x}^\top \boldsymbol{y} + \nu)^p \tag{1}$$

for some $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, where $\gamma, \nu \geq 0$ and $p \in \mathbb{N}$. Both parameters γ and ν can be absorbed by the input vectors by setting $\tilde{\boldsymbol{x}} := (\sqrt{\gamma}\boldsymbol{x}^\top, \sqrt{\nu})^\top \in \mathbb{R}^{d+1}$ and $\tilde{\boldsymbol{y}} := (\sqrt{\gamma}\boldsymbol{y}^\top, \sqrt{\nu})^\top \in \mathbb{R}^{d+1}$. Therefore, without loss of generality, we assume the kernel to be *homogeneous*, i.e., it can be written as

$$(\gamma \boldsymbol{x}^\top \boldsymbol{y} + \nu)^p = (\tilde{\boldsymbol{x}}^\top \tilde{\boldsymbol{y}})^p = (\tilde{\boldsymbol{x}}^{\otimes p})^\top \tilde{\boldsymbol{y}}^{\otimes p}. \tag{2}$$

Although its feature maps $\tilde{\boldsymbol{x}}^{\otimes p}, \tilde{\boldsymbol{y}}^{\otimes p}$ can be computed explicitly, they are $(d+1)^p$-dimensional and therefore infeasible to construct when d or p are large. For n data points, applying the kernel trick costs at least $\mathcal{O}(n^2)$ and is not possible when n is large. This makes randomized sketching, i.e., reducing the dimensionality of $\tilde{\boldsymbol{x}}^{\otimes p}$ and $\tilde{\boldsymbol{y}}^{\otimes p}$ through linear random projections, an attractive choice.

2.1 Sketching Tensor Products

We study sketches of tensor products $\otimes_{i=1}^p \boldsymbol{x}_i$ for some $\boldsymbol{x}_1 \in \mathbb{R}^{d_1}, \ldots, \boldsymbol{x}_p \in \mathbb{R}^{d_p}$. There exist several sketching techniques for this purpose (see Section 5). Here we focus on the following construction.

We generate $p \times D$ i.i.d. random weights $\boldsymbol{w}_{i,\ell} \in \mathbb{C}^{d_i}$ satisfying $\mathbb{E}[\boldsymbol{w}_{i,\ell}\overline{\boldsymbol{w}_{i,\ell}}^\top] = \boldsymbol{I}_{d_i}$ for $i \in \{1, \ldots, p\}, \ell \in \{1, \ldots, D\}$, where $\boldsymbol{I}_{d_i}$ is the identity matrix of size d_i. E.g., $\boldsymbol{w}_{i,\ell}$ can be a (complex) Rademacher vector or be sampled from the (complex) standard normal distribution.

We define a sketch $\boldsymbol{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D)^\top \in \mathbb{C}^{D \times d_1 \cdots d_p}$ with $\boldsymbol{s}_\ell = \otimes_{i=1}^p \boldsymbol{w}_{i,\ell}/\sqrt{D}$. A naive computation of $\boldsymbol{S}(\otimes_{i=1}^p \boldsymbol{x}_i)$ would cost $\mathcal{O}(D \prod_{i=1}^p d_i)$ time and memory, but we can exploit the following property of the tensor product:

$$(\otimes_{i=1}^p \boldsymbol{w}_{i,\ell})^\top (\otimes_{i=1}^p \boldsymbol{x}_i) = \prod_{i=1}^p \boldsymbol{w}_{i,\ell}^\top \boldsymbol{x}_i \tag{3}$$

that lets us compute $\boldsymbol{S}(\otimes_{i=1}^p \boldsymbol{x}_i)$ in $\mathcal{O}(D \sum_{i=1}^p d_i)$ using the r.h.s. of Eq. 3. In particular, $\otimes_{i=1}^p \boldsymbol{x}_i$ never needs to be constructed explicitly in this case.

Although our sketches are applicable to arbitrary tensor products, in this work we focus on feature maps of the polynomial kernel. That is, we set $\boldsymbol{x} = \boldsymbol{x}_1 = \cdots = \boldsymbol{x}_p$, such that $\otimes_{i=1}^p \boldsymbol{x}_i = \boldsymbol{x}^{\otimes p}$. For two inputs $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we define the approximate kernel $\hat{k}(\boldsymbol{x}, \boldsymbol{y}) := (\boldsymbol{S}\boldsymbol{x}^{\otimes p})^\top (\overline{\boldsymbol{S}\boldsymbol{y}^{\otimes p}})$, which is unbiased because

$$\mathbb{E}\left[\hat{k}(\boldsymbol{x}, \boldsymbol{y})\right] = \frac{1}{D} \sum_{\ell=1}^{D} \prod_{i=1}^{p} \boldsymbol{x}^\top \mathbb{E}[\boldsymbol{w}_{i,\ell}\overline{\boldsymbol{w}_{i,\ell}}^\top] \boldsymbol{y} = (\boldsymbol{x}^\top \boldsymbol{y})^p.$$

In this case, we may alternatively call $\Phi(\boldsymbol{x}) := \boldsymbol{S}\boldsymbol{x}^{\otimes p}$ a random feature map, which we express as

$$\Phi(\boldsymbol{x}) = (\boldsymbol{W}_1 \boldsymbol{x} \odot \cdots \odot \boldsymbol{W}_p \boldsymbol{x})/\sqrt{D}, \tag{4}$$

where $\boldsymbol{W}_i := (\boldsymbol{w}_{i,1}, \ldots, \boldsymbol{w}_{i,D})^\top$, to simplify the notation.

[1] https://github.com/joneswack/dp-rfs

Algorithm 1: Complex-to-Real (CtR) Sketches

Input: Data point $\boldsymbol{x} \in \mathbb{R}^d$
Choose dimension $D = 2k$ ($k \in \mathbb{N}$), degree $p \in \mathbb{N}$
Sample $\{\boldsymbol{W}_i\}_{i=1}^p$ with $\boldsymbol{W}_i \in \mathbb{C}^{D/2 \times d}$ independently according to one of the following sketch distributions:

- Gaussian: $(\boldsymbol{W}_i)_{\ell,k} \overset{i.i.d.}{\sim} \mathcal{CN}(0,1)$
- Rademacher: $(\boldsymbol{W}_i)_{\ell,k} \overset{i.i.d.}{\sim} \mathrm{Unif}(\{1,-1,\mathrm{i},-\mathrm{i}\})$
- ProductSRHT: $\boldsymbol{W}_i = \boldsymbol{P}_i \boldsymbol{H} \boldsymbol{D}_i$ (see Appendix 4)

Compute $\Phi_{\mathrm{C}}(\boldsymbol{x}) := \sqrt{2/D}\,(\boldsymbol{W}_1\boldsymbol{x} \odot \cdots \odot \boldsymbol{W}_p\boldsymbol{x})$
Return:
$\Phi_{\mathrm{CtR}}(\boldsymbol{x}) := (\mathrm{Re}\{\Phi_{\mathrm{C}}(\boldsymbol{x})_1\}, \ldots, \mathrm{Re}\{\Phi_{\mathrm{C}}(\boldsymbol{x})_{D/2}\}, \mathrm{Im}\{\Phi_{\mathrm{C}}(\boldsymbol{x})_1\}, \ldots, \mathrm{Im}\{\Phi_{\mathrm{C}}(\boldsymbol{x})_{D/2}\})^\top \in \mathbb{R}^D$

The random feature map (4) has originally been proposed by Kar & Karnick (2012) and been further studied in Hamid et al. (2014); Meister et al. (2019); Ahle et al. (2020) for the case of real-valued $\{\boldsymbol{W}_i\}_{i=1}^p$. Recently, Wacker et al. (2022, Thm. 3.1) derived a variance lower bound for $\hat{k}(\boldsymbol{x}, \boldsymbol{y})$, which can be obtained through Rademacher weights. They further showed that lower variances can be achieved using more general complex-valued $\{\boldsymbol{W}_i\}_{i=1}^p$ that subsume the real-valued case (Wacker et al., 2022, Thm. 3.3). Hereafter, we use Φ_{R} to denote a real-valued and Φ_{C} to denote a complex-valued random feature map, thus emphasizing their difference. The caveat of using Φ_{C} is that it requires the downstream model to handle complex data, which may incur additional computational costs.

The purpose of this work instead, is to analyze the real-valued kernel estimate $\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y}) := \mathrm{Re}\{\hat{k}(\boldsymbol{x}, \boldsymbol{y})\}$, which can be written as

$$\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{Re}\{\Phi_{\mathrm{C}}(\boldsymbol{x})\}^\top \mathrm{Re}\{\Phi_{\mathrm{C}}(\boldsymbol{y})\} + \mathrm{Im}\{\Phi_{\mathrm{C}}(\boldsymbol{x})\}^\top \mathrm{Im}\{\Phi_{\mathrm{C}}(\boldsymbol{y})\} = \Phi_{\mathrm{CtR}}(\boldsymbol{x})^\top \Phi_{\mathrm{CtR}}(\boldsymbol{y}), \tag{5}$$

where we call $\Phi_{\mathrm{CtR}}(\boldsymbol{x})$ a *Complex-to-Real (CtR)* sketch. Since it is real-valued, it can be used as a drop-in replacement for any input to a downstream model. The downside of CtR-sketches is that they are $2D$-dimensional. In order to yield a fair comparison with real sketches, we reduce the dimension of CtR-sketches to D by using half the number of rows for $\{\boldsymbol{W}_i\}_{i=1}^p$ from now onward. We summarize the construction of CtR-sketches in Alg. 1.

3 ANALYSIS OF CtR-SKETCHES

The following section is dedicated to the theoretical analysis of CtR-sketches for tensor products and feature maps of the polynomial kernel. For the first part of our analysis, we treat them as *linear* sketches in a high-dimensional tensor-product space (see Section 2). In the second part, we focus on the variances of CtR-sketches for the particular case of feature maps of the polynomial kernel in order to obtain useful insights for their practical application.

3.1 Concentration Bounds for Complex Sketches

We start by analyzing the complex sketch $\boldsymbol{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D)^\top \in \mathbb{C}^{D \times d_1 \cdots d_p}$ (see Section 2). Recall that $\boldsymbol{s}_\ell = \otimes_{i=1}^p \boldsymbol{w}_{i,\ell}/\sqrt{D}$ with $\boldsymbol{w}_{i,\ell} \in \mathbb{C}^{d_i}$ i.i.d. $\boldsymbol{x} \in \mathbb{R}^{d_1 \cdots d_p}$ can take on any value that may not necessarily result from a tensor product in our analysis. This makes our results more general and is required to derive the spectral guarantee (6) in Section 3.2. Moreover, we only study (complex) Gaussian/Rademacher distributions for $\boldsymbol{w}_{i,\ell}$ here since Rademacher distributions achieve a variance lower bound for the sketch in Eq. 4 as we show later in Thm. 3.4.

The following key lemma shows that $\boldsymbol{s}_\ell^\top \boldsymbol{x}$ has lower absolute moments if $\boldsymbol{s}_\ell$ is sampled from a complex Gaussian/Rademacher distribution instead of a real one. It is an extension of Ahle et al. (2020, Lem. 19) to complex $\boldsymbol{s}_\ell$.

Lemma 3.1 (Absolute Moment Bound) *Let $t \geq 2, p \in \mathbb{N}$, $C_t > 0$, $\boldsymbol{x} \in \mathbb{R}^{d_1 \cdots d_p}$ and $\boldsymbol{w}_i \in \mathbb{C}^{d_i}$ for $i = 1, \ldots, p$. If $\|\boldsymbol{w}_i^\top \boldsymbol{a}\|_{L^t} \leq C_t \|\boldsymbol{a}\|_2$ for all $\boldsymbol{a} \in \mathbb{R}^{d_i}$ and $\{\boldsymbol{w}_i\}_{i=1}^p$, then*

$$\|(\otimes_{i=1}^p \boldsymbol{w}_i)^\top \boldsymbol{x}\|_{L^t} \leq C_t^p \|\boldsymbol{x}\|_2 \quad \textit{holds.}$$

In particular, for $t = 2k$ with $k \in \mathbb{N}$, we obtain:

$$C_t = \sqrt{2}\pi^{-1/(2t)}\Gamma((t+1)/2)^{1/t} \qquad \textit{(real Gauss./Rad.)}$$
$$C_t = \Gamma(t/2+1)^{1/t} \qquad \textit{(complex Gauss./Rad.)}$$

which are tight constants. $\Gamma(\cdot)$ is the Gamma function.

Proof Appendix A.1, where we also bound C_t if $t \neq 2k$. ■

The left plot of Fig. 1 shows the constants C_t for different values of t and it becomes clear that higher order moments for the complex Gaussian/Rademacher distribution are smaller than for the real-valued one, with an increasing gain for larger t. This effect is again amplified with a larger p that enters the moment bounds exponentially.

The following theorem shows that complex sketches $\boldsymbol{S}\boldsymbol{x}$ thus require a lower sketching dimension D than real ones to preserve the norm of $\boldsymbol{x}$, which is a direct consequence of the tighter moment bounds in Lem. 3.1.

Theorem 3.2 (Norm Preservation) *Let $0 < \epsilon, 0 < \delta < \exp(-2)$, $\boldsymbol{x} \in \mathbb{R}^{d_1 \cdots d_p}$, $\boldsymbol{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D)^\top \in \mathbb{C}^{D \times d_1 \cdots d_p}$ with $\boldsymbol{s}_\ell = \otimes_{i=1}^p \boldsymbol{w}_{i,\ell}/\sqrt{D}$ and $\boldsymbol{w}_{i,\ell} \in \mathbb{C}^{d_i}$ be i.i.d. Gaussian/Rademacher samples. In order to guarantee*

$$\Pr\left\{ \left| \|\boldsymbol{S}\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2 \right| \leq \epsilon \|\boldsymbol{x}\|_2^2 \right\} \geq 1 - \delta, \quad \textit{we need}$$

$D = \mathcal{O}(\max\{C_4^{4p} \log(1/\delta)\epsilon^{-2}, (C_4^2 e/2)^p \log^p(1/\delta)\epsilon^{-1}\})$, *where C_4 is defined in Lem. 3.1 for the real/complex case.*

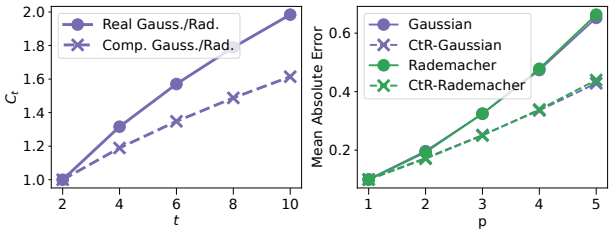


Figure 1: (Left) C_t over $t = 2k, k \in \mathbb{N}$. (Right) Mean $|\,\|\boldsymbol{S}\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2\,|$ over $3 \cdot 10^4$ samples of $\boldsymbol{S}$ for real/CtR-sketches with an equal number of 128 rows and $\boldsymbol{x} = \boldsymbol{a}^{\otimes p}, \boldsymbol{a} = (1/\sqrt{d}, \dots, 1/\sqrt{d})^\top \in \mathbb{R}^d, d = 64$.

Proof Appendix A.2, where we provide an additional bound for the case when $\delta \in (0, 1)$. ■

The upper bound on D in Thm. 3.2 is hence controlled by $C_4^{4p} = 3^p(2^p)$ and $C_4^{2p} = \sqrt{3}^p(\sqrt{2}^p)$ for real (complex) Gaussian/Rademacher sketches leading to a sharper dependence on p for the complex case. In particular, the 3^p dependence is tight for the real case as shown in the lower bound on D in Ahle et al. (2020, Appendix A.1), which makes our bound a remarkable improvement. It thus takes us one step closer to reaching the optimal $D = \Theta(\log(1/\delta)\epsilon^{-2})$ for Johnson-Lindenstrauss embeddings (Larsen & Nelson, 2017) that is independent from p, but has a prohibitive $\mathcal{O}(D\prod_{i=1}^p d_i)$ computational cost. Lastly, our result improves over Wacker et al. (2022, Thm. 3.4) that bounds errors relative to the L1-norm instead of the L2-norm, which makes their bound much looser than ours as explained in Appendix A.3.

3.2 Concentration Bounds for CtR-Sketches

It is easy to see that CtR-sketches directly inherit the guarantees in Thm. 3.2. Following the construction of CtR-features in Eq. 5, we define the $2D$-dimensional CtR-sketch

$$\boldsymbol{S}_{\mathrm{CtR}} := (\mathrm{Re}\{\boldsymbol{s}_1\}, \dots, \mathrm{Re}\{\boldsymbol{s}_D\}, \mathrm{Im}\{\boldsymbol{s}_1\}, \dots, \mathrm{Im}\{\boldsymbol{s}_D\})^\top$$

giving $\|\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{x}\|_2^2 = \sum_{\ell=1}^D \mathrm{Re}\{\boldsymbol{s}_\ell^\top\boldsymbol{x}\}^2 + \mathrm{Im}\{\boldsymbol{s}_\ell^\top\boldsymbol{x}\}^2 = \|\boldsymbol{S}\boldsymbol{x}\|_2^2$. We can thus substitute $\boldsymbol{S}$ in Thm. 3.2 by $\boldsymbol{S}_{\mathrm{CtR}}$ to obtain the same guarantees. For a fair comparison, we need to multiply the required number of features D in Thm. 3.2 by two when using the *same* number of rows for $\boldsymbol{S}_{\mathrm{CtR}}$ and $\boldsymbol{S}$. Crucially however, the improved dependence on p remains the same implying that CtR-sketches must outperform real-valued analogs when p is large enough. The right plot of Fig. 1 shows that this is already the case from $p \geq 2$ with a larger gain for larger p. A more detailed variance comparison follows in Section 3.3.

The following corollary of Thm. 3.2 shows that inner products as well as matrix products are preserved under the same conditions provided in the theorem.

Corollary 3.3 (Approximate Matrix Product) *Let* $0 < \epsilon, 0 < \delta < \exp(-2), \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{d_1 \cdots d_p}$ *and* $\boldsymbol{S}_{\mathrm{CtR}}$ *defined as in Section 3.2. In order to guarantee*

$$\Pr\left\{|(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{x})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{y}) - \boldsymbol{x}^\top\boldsymbol{y}| \leq \epsilon\,\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|\right\} \geq 1 - \delta,$$

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

$$\Pr\left\{\frac{\|(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{X})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{Y}) - \boldsymbol{X}^\top\boldsymbol{Y}\|_F}{\|\boldsymbol{X}\|_F\,\|\boldsymbol{Y}\|_F} \leq \epsilon\right\} \geq 1 - \delta,$$

$\boldsymbol{S}_{\mathrm{CtR}}$ *needs to have* $2D$ *rows with* D *being the same as in Thm. 3.2.*

Proof Appendix A.4. ■

In particular, Cor. 3.3 gives guarantees on the approximation error of polynomial kernels using CtR-sketches. In this case, we simply set $\boldsymbol{X} = \boldsymbol{A}^{\otimes p}, \boldsymbol{Y} = \boldsymbol{B}^{\otimes p}$, with $\boldsymbol{A}^{\otimes p}, \boldsymbol{B}^{\otimes p}$ being matrices whose columns are the polynomial kernel feature maps of some data points $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\{\boldsymbol{b}_i\}_{i=1}^m$, respectively, where $\boldsymbol{a}_i, \boldsymbol{b}_i \in \mathbb{R}^d$.

We can directly derive spectral kernel approximation guarantees from the approximate matrix product property as shown in Appendix A.5. Let $\boldsymbol{K} := (\boldsymbol{A}^{\otimes p})^\top\boldsymbol{A}^{\otimes p} \in \mathbb{R}^{n\times n}$ be the gram matrix for the points $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\hat{\boldsymbol{K}} := (\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{A}^{\otimes p})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{A}^{\otimes p})$ be its randomized approximation. Then with probability at least $1 - \delta$, we have

$$(1 - \epsilon)(\boldsymbol{K} + \lambda\boldsymbol{I}) \preceq \hat{\boldsymbol{K}} + \lambda\boldsymbol{I} \preceq (1 + \epsilon)(\boldsymbol{K} + \lambda\boldsymbol{I}) \quad (6)$$

for some $\lambda \geq 0$, if $\boldsymbol{S}_{\mathrm{CtR}}$ has $2Ds_\lambda(\boldsymbol{K})^2$ rows with D being the same as in Thm. 3.2 and $s_\lambda(\boldsymbol{K}) = \mathrm{Tr}\{\boldsymbol{K}(\boldsymbol{K} + \lambda\boldsymbol{I})^{-1}\} \leq n$ being the λ-statistical dimension of $\boldsymbol{K}$.

The spectral approximation guarantee directly implies statistical guarantees for downstream kernel-based learning applications, such as bounds on the empirical risk of kernel ridge regression (Avron et al., 2017, Lem. 2). The quadratic dependence on $s_\lambda(\boldsymbol{K})$ is not optimal and arises due to the element-wise error bound in Cor. 3.3. A linear dependence could be achieved by bounding the operator norm instead as it is done in Ahle et al. (2020, Section 5). As the focus of this work is to obtain a sharp dependence w.r.t. p and δ, we leave this issue to future work and focus on a careful variance analysis of CtR-sketches instead.

3.3 Variances of CtR-Sketches for Polynomial Kernels

In this section, we derive the closed form variances of CtR-sketches for the specific task of polynomial kernel approximation, and compare them against their real-valued analogs. This analysis is crucial, since we saw in Thm. 3.2 that the improvement of CtR over real-valued sketches is because of a lower fourth moment C_4 as defined in Lem. 3.1. To be more precise, let $\boldsymbol{s}_\ell^\top\boldsymbol{x}$ be a single element of our complex sketch $\boldsymbol{S}\boldsymbol{x} \in \mathbb{C}^D$ as defined in Section 3.1. Then we have $\mathbb{E}[|\boldsymbol{s}_\ell^\top\boldsymbol{x}|^4] \leq (C_4\,\|\boldsymbol{x}\|_2/\sqrt{D})^4$ as

implied by Lem. 3.1. This is equal to the second moment $\mathbb{E}[|\boldsymbol{s}_\ell^\top \boldsymbol{x}\overline{\boldsymbol{s}_\ell^\top \boldsymbol{y}}|^2]$ for two inputs $\boldsymbol{x}, \boldsymbol{y}$ when $\boldsymbol{x} = \boldsymbol{y}$, which in turn is directly linked to the variance of $\boldsymbol{s}_\ell^\top \boldsymbol{x}\overline{\boldsymbol{s}_\ell^\top \boldsymbol{y}}$ via

$$\mathbb{E}\left[\left|\boldsymbol{s}_\ell^\top \boldsymbol{x}\overline{\boldsymbol{s}_\ell^\top \boldsymbol{y}}\right|^2\right] = \mathbb{V}\left[\boldsymbol{s}_\ell^\top \boldsymbol{x}\overline{\boldsymbol{s}_\ell^\top \boldsymbol{y}}\right] + \frac{1}{D^2}(\boldsymbol{x}^\top \boldsymbol{y})^2.$$

The purpose of this section is therefore to carry out a careful variance analysis, elucidating conditions on $\boldsymbol{x}$ and $\boldsymbol{y}$ under which CtR-sketches perform better than real-valued analogs *in practice*, and *beyond the worst-case scenario*: $\boldsymbol{x} = \boldsymbol{y}$ as explained in Appendix A.4.

As we focus on polynomial kernels from now onward, we restrict our input space to vectors $\boldsymbol{x}^{\otimes p}, \boldsymbol{y}^{\otimes p} \in \mathbb{R}^{d^p}$ for some $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. In this case, we can write:

$$(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{x}^{\otimes p})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{y}^{\otimes p}) = \hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y}) \quad \text{(as in Eq. 5)}$$

Hence, the approximate kernel and its variance $\mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})]$ depend directly on $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. Let further $\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{S}\boldsymbol{x}^{\otimes p})^\top(\overline{\boldsymbol{S}\boldsymbol{y}^{\otimes p}})$. In Appendix B.1, we show that CtR-sketches have the following variance structure:

$$\mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})] = \frac{1}{2}\left(\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] + \mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]\right) \quad (7)$$
$$\text{with} \quad \mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] := \mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})^2] - (\boldsymbol{x}^\top \boldsymbol{y})^{2p}$$

being the *pseudo-variance* of $\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})$ and $\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]$ its variance.

Our major contribution of this section is to derive $\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]$ and $\mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]$ for Gaussian/Rademacher sketches in Section B.2 and we summarize these results in Table 1. For a direct comparison, we also add the variances of real sketches Φ_{R} (see Section 2.1) to Table 1. The question that we address in the following is: Does the CtR estimator in Eq. 5 yield lower variances than $\hat{k}_{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y}) = \Phi_{\mathrm{R}}(\boldsymbol{x})^\top \Phi_{\mathrm{R}}(\boldsymbol{y})$ if Φ_{CtR} and Φ_{R} have the *same* output dimension D? We show next that this is indeed the case.

3.4 Variance Reduction of CtR-Sketches

We begin by studying the variance reduction properties of Gaussian/Rademacher CtR-sketches over their real-valued analogs. Let $\Phi_{\mathrm{R}} : \mathbb{R}^d \to \mathbb{R}^D$ be a real-valued sketch (see Section 2.1) and $\Phi_{\mathrm{CtR}} : \mathbb{R}^d \to \mathbb{R}^D$ a CtR-sketch as defined in Alg. 1. Let $\hat{k}_{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y})$ and $\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})$ (5) be the respective approximate kernels for some $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. Then we can provide the following theorem for Rademacher sketches.

Theorem 3.4 (CtR-Rademacher advantage) *Let* $a = \sum_{i \neq j'}^d x_i x_{j'} y_i y_{j'}$, $b_j = (\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|)^{2j} - (\sum_i x_i^2 y_i^2)^j \geq 0$. *Then* $\mathbb{V}[\hat{k}_{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y})] - \mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})]$ *is equal to*

$$\frac{1}{D}\sum_{k=2}^{p}\sum_{j=0}^{k-1}\binom{p}{k}\binom{k}{j} b_j\, a^{p-j} \geq 0 \quad \textit{if} \quad a \geq 0.$$

Furthermore, if $a \geq 0$, *CtR-Rademacher sketches achieve the lowest possible variance for* $\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})$ *(5) assuming the entries of* $\{\boldsymbol{W}_i\}_{i=1}^p$ *in Eq. 4 are i.i.d. If* $a < 0$, *the lowest possible variance is attained by real Rademacher sketches instead, i.e., using* $\hat{k}_{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y})$.

Proof The variance reduction and lowest variance property are proved in Appendix B.3.2 and B.2, respectively. ■

The theorem tells us that Φ_{CtR} should be preferred over Φ_{R} when $a \geq 0$ for two given inputs $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and the variance gap increases as p increases. The condition

$$a = \sum_{i=1}^d \sum_{j' \neq i}^d x_i x_{j'} y_i y_{j'} = (\boldsymbol{x}^\top \boldsymbol{y})^2 - \sum_{i=1}^d x_i^2 y_i^2 \geq 0$$

always holds if $\boldsymbol{x}, \boldsymbol{y}$ are non-negative or if they are parallel, thus leading to improved worst-case guarantees in Thm. 3.2 and Cor. 3.3. Non-negative data typically appears in applications for polynomial kernels such as categorical and image data, as well as outputs of convolutional neural networks. We carry out corresponding numerical experiments in Section 6. We also note that CtR-Rademacher sketches outperform real-valued analogs when the condition $a \geq 0$ is not always met as shown in Appendix D.1. This is because $a \geq 0$ always holds for the diagonal elements of the kernel matrix, leading to an inherent bias towards $a \geq 0$.

We can additionally provide the following theorem for Gaussian sketches proved in Appendix B.3.1.

Theorem 3.5 (CtR-Gaussian advantage) *For any* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\mathbb{V}[\hat{k}_{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y})] - \mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})]$ *is equal to*

$$\frac{1}{D}\sum_{k=0}^{p-1}\binom{p}{k}(2^k - 1)(\boldsymbol{x}^\top \boldsymbol{y})^{2k}\left(\|\boldsymbol{x}\|^2\,\|\boldsymbol{y}\|^2\right)^{p-k} \geq 0.$$

Thus, regardless of the input data, Φ_{CtR} should be preferred over Φ_{R} when using Gaussian sketches. The advantage again increases with p.

4 ProductSRHT

In this section, we propose a novel structured Rademacher sketch. Our sketch is called *ProductSRHT* and is closely related to TensorSRHT (Ahle et al., 2020, Def. 15). A major difference is that we are able to obtain the variance of ProductSRHT in closed form showing its statistical advantages over unstructured sketches. The variance derivation is contained in Appendix C.1. We also embed ProductSRHT into our CtR-framework and compare its variance against CtR-Rademacher sketches in Section 4.1.

Both ProductSRHT and TensorSRHT achieve a $\mathcal{O}(p(D + d\log d))$ runtime through structured Hadamard matrices that we introduce in the following. Let $n := 2^m$ with $m \in \mathbb{N}$, and $\boldsymbol{H}_n \in \{1, -1\}^{n \times n}$ be the unnormalized

Table 1: Variances of complex $\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})$ and real $\hat{k}_{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y})$ and pseudo-variances of complex $\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})$ are shown. $\mathbb{V}^{(p)}_{\mathrm{Rad.}}, \mathbb{V}^{(1)}_{\mathrm{Rad.}}$ and $\mathbb{PV}^{(p)}_{\mathrm{Rad.}}, \mathbb{PV}^{(1)}_{\mathrm{Rad.}}$ are the Rademacher variances / pseudo-variances for a given p and $p = 1$, respectively.

Sketch	**Variance** $\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]$ $(q = 1)$ and $\mathbb{V}[\hat{k}_{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y})]$ $(q = 2)$	**Pseudo-Variance** $\mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]$
Gaussian	$D^{-1}[(\Vert\boldsymbol{x}\Vert^2 \Vert\boldsymbol{y}\Vert^2 + q(\boldsymbol{x}^\top \boldsymbol{y})^2)^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p}]$	$D^{-1}[(2(\boldsymbol{x}^\top \boldsymbol{y})^2)^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p}]$
Rademacher	$D^{-1}[(\Vert\boldsymbol{x}\Vert^2 \Vert\boldsymbol{y}\Vert^2 + q((\boldsymbol{x}^\top \boldsymbol{y})^2 - \sum_{i=1}^d x_i^2 y_i^2))^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p}]$	$D^{-1}[(2(\boldsymbol{x}^\top \boldsymbol{y})^2 - \sum_{i=1}^d x_i^2 y_i^2)^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p}]$
ProductSRHT	$\mathbb{V}^{(p)}_{\mathrm{Rad.}} - (1 - 1/D) \cdot [(\boldsymbol{x}^\top \boldsymbol{y})^{2p} - (\mathrm{C}_{\mathrm{Var.}})^p]$	$\mathbb{PV}^{(p)}_{\mathrm{Rad.}} - (1 - 1/D) \cdot [(\boldsymbol{x}^\top \boldsymbol{y})^{2p} - (\mathrm{C}_{\mathrm{PVar.}})^p]$
	$\mathrm{C}_{\mathrm{Var.}} = (\boldsymbol{x}^\top \boldsymbol{y})^2 - (\lceil D/d \rceil d - 1)^{-1} \mathbb{V}^{(1)}_{\mathrm{Rad.}}$	$\mathrm{C}_{\mathrm{PVar.}} = (\boldsymbol{x}^\top \boldsymbol{y})^2 - (\lceil D/d \rceil d - 1)^{-1} \mathbb{PV}^{(1)}_{\mathrm{Rad.}}$

Hadamard matrix, which is recursively defined as

$$\boldsymbol{H}_{2n} := \begin{bmatrix} \boldsymbol{H}_n & \boldsymbol{H}_n \\ \boldsymbol{H}_n & -\boldsymbol{H}_n \end{bmatrix}, \quad \text{with} \quad \boldsymbol{H}_2 := \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

From now onward, we always use $\boldsymbol{H}_d \in \{1, -1\}^{d \times d}$ with d being the dimension of the input vectors, assuming $d = 2^m$ for some $m \in \mathbb{N}$. If $d \neq 2^m$ for any m, we pad the input vectors with 0 until their dimension becomes 2^m for some m. We have $\boldsymbol{H}_d \boldsymbol{H}_d^\top = \boldsymbol{H}_d^\top \boldsymbol{H}_d = d\boldsymbol{I}_d$ and the recursive definition of $\boldsymbol{H}_d$ gives rise to the Fast Walsh-Hadamard transform (Fino & Algazi, 1976) that multiplies $\boldsymbol{H}_d$ with a vector $\boldsymbol{a} \in \mathbb{R}^d$ in $\mathcal{O}(d \log d)$ instead of $\mathcal{O}(d^2)$ time, while the matrix $\boldsymbol{H}_d$ does not need to be stored in memory. We drop the subscript d from now for ease of presentation.

We describe (CtR-)ProductSRHT in Alg. 2. It uses structured matrices $\{\boldsymbol{W}_i\}_{i=1}^p$ in Eq. 4, which are formed through an element-wise multiplication of the rows of $\boldsymbol{H}$ with a Rademacher vector, imposing an orthogonality condition on these rows. This ultimately leads to a variance reduction that we analyze next. Finally, the rows of $\{\boldsymbol{W}_i\}_{i=1}^p$ are randomly up/downsampled to cover the case $D \neq d$.

4.1 Variance of CtR-ProductSRHT

A major contribution of this work is to derive the variance of our proposed CtR-ProductSRHT sketch in closed form, which requires the derivation of the variance and pseudo-variance of complex ProductSRHT as shown in Eq. 7. They are derived in Section C.1 and we summarize them in Table 1. We also derive the variance of real ProductSRHT in Section C.1.2 and add it to Table 1 for comparison.

ProductSRHT can yield lower variances than Rademacher sketches as it removes the i.i.d. constraint between the $\{\boldsymbol{w}_{i,\ell}\}_{\ell=1}^D$ in Eq. 3. In fact, these vectors are mutually orthogonal for two $\ell \neq \ell'$ when the $p_{i,\pi(\ell)}$-th and the $p_{i,\pi(\ell')}$-th column of $\boldsymbol{H}$ are distinct, since $\boldsymbol{H}$ has orthogonal rows and columns. This dependence introduces the term

$$\mathrm{R}_{\mathrm{Var./PVar.}} := (1 - 1/D)[(\boldsymbol{x}^\top \boldsymbol{y})^{2p} - (\mathrm{C}_{\mathrm{Var./PVar.}})^p]$$

that is subtracted from the original Rademacher variance $\mathbb{V}^{(p)}_{\mathrm{Rad.}}$ and pseudo-variance $\mathbb{PV}^{(p)}_{\mathrm{Rad}}$, respectively, as shown in Table 1, where we also define $\mathrm{C}_{\mathrm{Var.}}$ and $\mathrm{C}_{\mathrm{PVar.}}$.

If p is odd, $(\mathrm{C}_{\mathrm{Var.}})^p \leq (\boldsymbol{x}^\top \boldsymbol{y})^{2p}$ holds because $\mathbb{V}^{(1)}_{\mathrm{Rad.}} \geq 0$ and therefore $\mathrm{R}_{\mathrm{Var.}} \geq 0$ holds. In this case, the variance of complex/real ProductSRHT is upper-bounded by the complex/real Rademacher variance $\mathbb{V}^{(p)}_{\mathrm{Rad.}} \geq 0$.

If we further have $\mathbb{PV}^{(1)}_{\mathrm{Rad.}} = \sum_{i=1}^d \sum_{j \neq i}^d x_i x_j y_i y_j \geq 0$, the pseudo-variance of complex ProductSRHT is also upper-bounded by the Rademacher pseudo-variance. This is because $0 \leq (\mathrm{C}_{\mathrm{PVar.}})^p \leq (\boldsymbol{x}^\top \boldsymbol{y})^{2p}$ and $\mathrm{R}_{\mathrm{PVar.}} \geq 0$ hold. Note that this is exactly the same condition as $a \geq 0$ in Thm. 3.4. CtR-ProductSRHT thus has a lower pseudo-variance than CtR-Rademacher sketches exactly when CtR-Rademacher sketches are better than real ones.

As both the variance and the pseudo-variance of complex ProductSRHT are upper-bounded by the ones of complex Rademacher sketches under the above conditions, CtR-ProductSRHT is guaranteed to have a lower variance than CtR-Rademacher sketches through Eq. 7 in this case. Moreover, CtR-ProductSRHT inherits the variance reduction of CtR-Rademacher sketches over their real analogs because the Rademacher variance and pseudo-variance both enter the ones of complex ProductSRHT (see Table 1).

5 RELATED WORK

In this work, we study the sketches for tensor products presented in Section 2.1 building on previous works by Kar & Karnick (2012); Hamid et al. (2014); Meister et al. (2019); Ahle et al. (2020); Wacker et al. (2022). However, there exist alternatives that we have not mentioned so far.

Pham & Pagh (2013) have proposed **TensorSketch**, which is a convolution of CountSketches (Charikar et al., 2002). TensorSketch requires $D = \mathcal{O}(3^p s_\lambda(\boldsymbol{K})^2/(\delta\epsilon^2))$ to satisfy Eq. 6 (Avron et al., 2014) and thus has weaker guarantees w.r.t. δ and p than CtR-Gaussian/Rademacher sketches. There is also no closed form variance formula available for this sketch[2]. Yet, it achieves state-of-the-art performance in practice as we show in Section 6. It is also faster than Gaussian/Rademacher sketches taking only $\mathcal{O}(p(D \log D + d))$ instead of $\mathcal{O}(pdD)$ via the *Fast Fourier Transform*.

[2] Pham & Pagh (2013) contains a variance formula, but makes the simplifying assumption that TensorSketch has the same variance as CountSketch applied to tensorized inputs. Avron et al. (2014) conduct a more careful analysis to obtain an upper bound.

Algorithm 2: (CtR-) ProductSRHT

Input: Data point $\boldsymbol{x} \in \mathbb{R}^d$, projection dimension $D \in \mathbb{N}$
Pad $\boldsymbol{x}$ with zeros so that d becomes a power of 2, let $B = \lceil \frac{D}{d} \rceil$ be the number of stacked projection blocks
forall $i \in \{1, \dots, p\}$ **do**
  Generate a diagonal matrix $\boldsymbol{D}_i \in \mathbb{C}^{d \times d}$ with diagonal elements:
    $(\boldsymbol{D}_i)_{1,1}, \dots, (\boldsymbol{D}_i)_{d,d} \overset{i.i.d.}{\sim} \mathrm{Unif}(\{1, -1\})$ (real case)
    $(\boldsymbol{D}_i)_{1,1}, \dots, (\boldsymbol{D}_i)_{d,d} \overset{i.i.d.}{\sim} \mathrm{Unif}(\{1, -1, \mathrm{i}, -\mathrm{i}\})$ (complex case)
  Generate a random sampling matrix $\boldsymbol{P}_i \in \{1, 0\}^{D \times d}$ as follows:
    Let $\boldsymbol{p}_i = (p_{i,1}, \dots, p_{i,Bd})^\top \in \mathbb{R}^{Bd}$ be the B-times concatenation of $(1, \dots, d)^\top$
    Randomly permute the indices $1, \dots, Bd$ to $\pi(1), \dots, \pi(Bd)$
    Set $\boldsymbol{P}_i = (\boldsymbol{e}_{p_{i,\pi(1)}}, \dots, \boldsymbol{e}_{p_{i,\pi(D)}})^\top$, where $\boldsymbol{e}_{p_{i,\pi(\ell)}} \in \{1, 0\}^d$ is equal to 1 at position $p_{i,\pi(\ell)}$ and 0 elsewhere
  Set $\boldsymbol{W}_i = \boldsymbol{P}_i \boldsymbol{H} \boldsymbol{D}_i$
end
Return: $\Phi_{\mathrm{R}}(\boldsymbol{x})$ using Eq. 4 for ProductSRHT or $\Phi_{\mathrm{CtR}}(\boldsymbol{x})$ using Alg. 1 for CtR-ProductSRHT

Structured Rademacher sketches based on the *Subsampled Randomized Hadamard Transform (SRHT)* (Tropp, 2011) have been proposed by Hamid et al. (2014), and a similar sketch called **TensorSRHT** by Ahle et al. (2020), referring to the fact that SRHT is implicitly applied to a tensorized version of the input. Both sketches use the Fast Walsh-Hadamard Transform (Fino & Algazi, 1976) for faster projections. Our (CtR-) ProductSRHT sketch is closely related. Notably, both TensorSRHT and our sketch have a runtime of $\mathcal{O}(p(d \log d + D))$ and are thus faster than TensorSketch when $D > d$. Unlike previous works, we derive the variance for our ProductSRHT sketch in closed form, showing statistical advantages over Rademacher sketches.

Recent research has focused on *meta-algorithms* that aim to improve the approximation error of existing sketches (Hamid et al., 2014; Ahle et al., 2020; Song et al., 2021). In particular, Ahle et al. (2020) managed to reduce the exponential dependence of D on p to polynomial by using a hierarchical construction. The sketches proposed in this work are compatible with these methods and can serve as their base sketches. In fact, we combine the **hierarchical** construction by Ahle et al. (2020) and **CRAFT maps** by (Hamid et al., 2014) with CtR-sketches in Section 6.

A fundamentally different approach are **Spherical Random Features (SRF)** (Pennington et al., 2015) that require a preprocessing step and yield biased polynomial kernel approximations for data on the unit-sphere. SRF can only be applied to inhomogeneous polynomial kernels and work well for large p. We adapt our experiments in Section 6 accordingly to accommodate a comparison against SRF.

6 EXPERIMENTS

In this section, we carry out a systematic comparison of the CtR-sketches presented in this work against their real-valued analogs as well as TensorSketch and SRF. We also combine CtR-ProductSRHT and TensorSketch with Ahle et al. (2020, Alg. 1) denoted as Hierarchical TensorSketch/CtR-ProductSRHT. Moreover, we add CRAFT maps (Hamid et al., 2014) denoted as CRAFT TensorSketch/CtR-ProductSRHT to this comparison.

6.1 Experimental Setup

Data sets We use MNIST (Lecun et al., 1998), and convolutional features[3] for CIFAR-10 (Krizhevsky et al., 2009) and CUB-200 (Welinder et al., 2010) as our data sets for the evaluation in this section. All three data sets contain only non-negative inputs to ensure that the condition of Thm. 3.4 is met. Additional experiments with zero-centered data and more data sets are contained in Appendix D.

Target kernel and its approximation Except for Section 6.4, we follow Pennington et al. (2015) and restrict our experiments to the polynomial kernel

$$k(\boldsymbol{x}, \boldsymbol{y}) = \left(\left(1 - 2/a^2\right) + 2/a^2 \, \boldsymbol{x}^\top \boldsymbol{y} \right)^p, \quad \text{with} \quad a \geq 2,$$

with for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ having unit-norm, as this allows a comparison against SRF. In particular, we set $a = 2$ to assign the largest weight possible to high polynomial degrees in the binomial expansion of the kernel, thereby making its approximation more challenging. Further results for non unit-norm data are contained in Section 6.4. We denote by D the feature map dimension that we ensure to be equal between CtR- and non-CtR sketches. For CRAFT maps, the intermediate up-projection dimension is fixed to $E = 2^{15}$. We measure the kernel approximation quality through the relative Frobenius norm error, which is defined as $\|\hat{\boldsymbol{K}} - \boldsymbol{K}\|_F / \|\boldsymbol{K}\|_F$, where $\hat{\boldsymbol{K}}$ is the random feature approximation of the exact kernel matrix $\boldsymbol{K}$ evaluated on a subset of the test data of size 1000 that is resampled for each seed used in these experiments.

[3] For CIFAR-10 (CUB-200), we use convolutional outputs of a ResNet34 (He et al., 2016) (VGG-M (Chatfield et al., 2014)) pretrained on ImageNet (Russakovsky et al., 2015).

All time benchmarks are run on an NVIDIA P100 GPU and PyTorch 1.10 (Paszke et al., 2019) with CUDA 10.2.

6.2 Kernel Approximation and GP Classification

We start by comparing the sketches discussed in this work for the downstream task of Gaussian Process (GP) classification. We model GP classification as a multi-class GP regression problem with transformed labels (Milios et al., 2018), for which we obtain closed-form solutions to measure the effects of the random feature approximations in isolation without the need for convergence verification.

Fig. 2 shows the result of this comparison. CtR-sketches generally result in lower kernel approximation errors than their real-valued analogs, with an increased effect for a larger degree $p = 7$. Overall, we see that ProductSRHT performs better than Gaussian and Rademacher sketches, and comparable to TensorSketch. The hierarchical extension of CtR-ProductSRHT/TensorSketch only improves results for $p = 7$, and performs worse for $p = 3$. CRAFT maps on the other hand always improve results.

Although similar trends can be observed for test errors, differences between the methods become only strongly noticeable for large $p = 7$. This makes sense, since all sketches become less optimal for larger p, amplifying their difference in statistical efficiency. For $D = 2^9$ and $p = 7$, CtR-sketches yield around 5% / 2.5% / 1% improvement for Gaussian / Rademacher / ProductSRHT, respectively. The absolute error difference among all methods decreases for larger D, but their relative improvement remains.

Kernel approximations for SRF are generally biased with a decreasing bias for larger p (Pennington et al., 2015, Section 4). We thus see that the relative Frobenius norm error for SRF stagnates for $p = 3$ when D is large, while the one for the other sketches continues decreasing. Test errors for SRF are also worse in this case. For $p = 7$, SRF kernel approximation errors tend to be lower than for CtR-ProductSRHT/TensorSketch and are comparable to their CRAFT extensions (slightly worse for MNIST, slightly better for CIFAR-10). SRF test errors are slightly worse than for the CRAFT sketches. In summary, CRAFT CtR-ProductSRHT and CRAFT TensorSketch yield the lowest test errors and kernel approximation errors (except for CIFAR-10 and $p = 7$, where SRF has lower errors).

6.3 Feature Construction Time Comparison

In the following, we carry out a feature construction time comparison of the methods presented in this work against TensorSketch that has a time complexity of $\mathcal{O}(p(D \log D + D))$ and SRF. Recall that our proposed ProductSRHT approach in Section 4 has a time complexity of $\mathcal{O}(p(d \log d + D))$ and is thus faster in theory when $D > d$. The left plot in Fig. 3 shows that this is also the case in practice.

Table 2: Projection time / downstream time ratio ($p = 3$).

D	SRF	Prod. SRHT	+ CtR	+ CtR + CRAFT	Tensor-Sketch	+ CRAFT
2^9	3.51	2.96	3.32	6.09	5.13	9.98
2^{11}	0.40	0.33	0.38	0.67	0.58	1.08
2^{13}	0.04	0.03	0.04	0.06	0.06	0.10

The construction times of real ProductSRHT and CtR-ProductSRHT have a smaller slope with respect to D than the other sketches leading to the lowest feature construction times together with SRF, in particular when $D \gg d$. There is a small computational overhead for CtR-ProductSRHT compared to ProductSRHT because CtR-ProductSRHT initially requires two Hadamard-projections (real and imaginary parts), but uses the same upsampling matrix leading to the same scaling property with respect to D. The right plot of Fig. 3 shows that SRF kernel approximations are strongly biased, making CtR-ProductSRHT the most accurate sketch for $p = 3$.

Faster feature construction times matter in practice. Although CRAFT maps enjoy a strong performance in Section 6.2, they can be expensive to compute due to the up-projection to $E = 2^{15}$ before down-projecting to D. Table 2 shows the ratio of feature construction time against solving the downstream GP model for MNIST. The ratio decays with larger D since solving the downstream model scales as $\mathcal{O}(nD^2 + D^3)$. For small D, the feature construction may dominate however. We also see that (CtR-) ProductSRHT is generally faster than TensorSketch. Moreover, feature construction times can heavily influence online learning scenarios in which the optimization algorithm requires a forward and backward pass through the feature map for every iteration.

6.4 Online Learning for Fine-Grained Recognition

We follow Gao et al. (2016) and carry out an online learning experiment using convolutional features from the CUB-200 data set (see Section 6.1). The task of fine-grained visual recognition is about the classification of pictures *within* their subordinate categories (200 bird species in this case). Here feature maps of low-degree polynomial kernels have proven very effective, but lead to classification layers with too many parameters due to high-dimensional inputs.

Gao et al. (2016) therefore use TensorSketch to reduce the dimension of explicit polynomial feature maps. We compare our methods against theirs and against SRF in Appendix D.3. (CtR-) ProductSRHT achieves the same test errors as TensorSketch, while being faster, especially when using CRAFT maps (almost 2x speedup). SRF is fast, but achieves only 75% test error compared to 30% achieved by the other methods. This is because SRF requires the unit-normalization of the convolutional features, hence loses important information. Polynomial kernels have thus im-

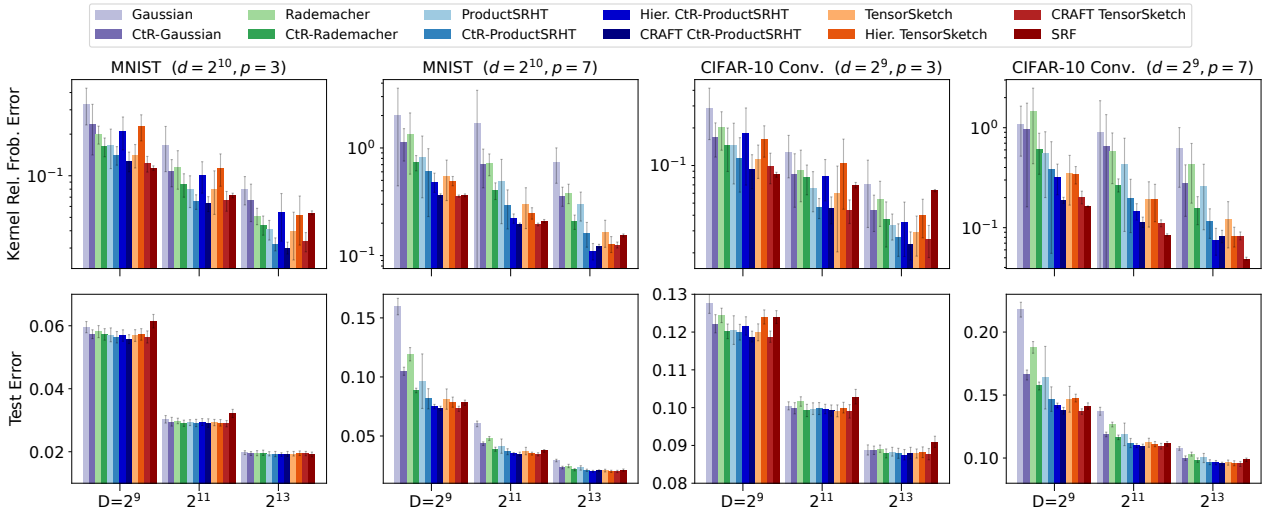


Figure 2: MNIST and CIFAR-10 comparison for $p = 3$ and $p = 7$ averaged over 20 seeds. Due to space limitations, we only show results for $D \in \{2^9, 2^{11}, 2^{13}\}$. Results for $D \in \{2^i\}_{i=8}^{13}$ and more data sets are contained in Appendix D.

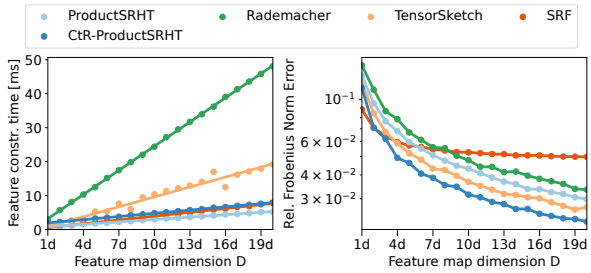


Figure 3: (Left) Feature construction time, (right) kernel approximation error, against feature map dimension D for $p = 3$ on 1000 random MNIST samples.

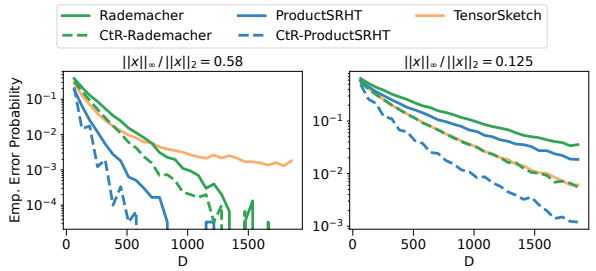


Figure 4: Emp. $\Pr\{|\|\boldsymbol{S}\boldsymbol{x}^{\otimes p}\|_2^2 - \|\boldsymbol{x}^{\otimes p}\|_2^2| \geq \epsilon \|\boldsymbol{x}^{\otimes p}\|_2^2\}$ for $\epsilon = 0.25$, $d = 64$, $p = 2$. (Left) $\boldsymbol{x} = (\sqrt{d}, \sqrt{d}, 1, \ldots, 1)^\top$, $r = 0.58$; (Right) $\boldsymbol{x} = (1, \ldots, 1)^\top$, $r = 0.125$.

portant applications *beyond unit-normalized* data, which is neglected in Pennington et al. (2015).

6.5 Error Bound Comparison

Lastly, we compare the empirical error probability of (CtR-) Rademacher/ProductSRHT against TensorSketch[4] for two fixed vectors $\boldsymbol{x} \in \mathbb{R}^d$ with a different maximum-to-norm-ratio $r := \|\boldsymbol{x}\|_\infty / \|\boldsymbol{x}\|_2$. TensorSketch can be seen as a CountSketch (Charikar et al., 2002) in a tensorized vector space (Pham & Pagh, 2013). Weinberger et al. (2009, Thm. 3) show that the error probability of CountSketches is heavily influenced by r due to hashing collisions.

This is also the case for TensorSketch as shown in Fig. 4, i.e., it converges much slower for $r = 0.58$ with an empirical error probability that is two orders of magnitude larger than for our methods for large D. For an extended discussion, see Meister et al. (2019, Section 4.1).

[4]We did not add SRF to this comparison because $\|\Phi(\boldsymbol{x})\|_2^2 = \Phi(\boldsymbol{x})^\top \Phi(\boldsymbol{x}) = \frac{1}{D} \sum_{\ell=1}^{D} \cos(\boldsymbol{\omega}_\ell^\top (\boldsymbol{x} - \boldsymbol{x})) = 1$ has zero variance.

7 CONCLUSION

The goal of research on random projections for tensor products is to achieve the optimal Johnson-Lindenstrauss embedding dimension of $D = \Theta(\epsilon^{-2} \log(1/\delta))$, i.e., without dependence on p, and without high computational costs. A recent work by Ahle et al. (2020) has improved the unwanted exponential dependence of $D = \mathcal{O}(3^p)$ to polynomial by using a hierarchical construction of well-known base sketches. However, we showed empirically in Section 6 that their method only yields improvements for large p and yields worse performance for small p.

In this work, we took a different approach by modifying the base sketch sampling distribution directly, thus achieving $D = \mathcal{O}(2^p)$ instead of $D = \mathcal{O}(3^p)$. Although still not being optimal, our method already leads to improvements from $p \geq 2$ and can be combined with other meta-algorithms. Moreover, we achieved state-of-the-art results in terms of accuracy *and* speed in our experiments. We thus uncovered an exciting angle of improvement that can be further leveraged in future research.

Acknowledgements

We thank Motonobu Kanagawa for helpful discussions. MF gratefully acknowledges support from the AXA Research Fund and the Agence Nationale de la Recherche (grant ANR-18-CE46-0002 and ANR-19-P3IA-0002). RO started this work while interning at the Criteo AI lab in Paris.

References

Ahle, T. D., Kapralov, M., Knudsen, J. B. T., Pagh, R., Velingker, A., Woodruff, D. P., and Zandieh, A. Oblivious sketching of high-degree polynomial kernels. In *Proceedings of the Thirty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 141–160. Society for Industrial and Applied Mathematics, 2020.

Aschard, H. A perspective on interaction effects in genetic association studies. *Genetic epidemiology*, 40(8): 678–688, 2016.

Avron, H., Nguyen, H. L., and Woodruff, D. P. Subspace embeddings for the polynomial kernel. In *Advances in Neural Information Processing Systems 27*, pp. 2258–2266. Curran Associates, Inc., 2014.

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 253–262. PMLR, 2017.

Charikar, M., Chen, K., and Farach-Colton, M. Finding frequent items in data streams. In *Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, pp. 693–703. Springer-Verlag, 2002.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*. BMVA Press, 2014.

Choromanski, K., Rowland, M., and Weller, A. The unreasonable effectiveness of structured random orthogonal embeddings. In *Advances in Neural Information Processing Systems 31*, pp. 218–227. Curran Associates Inc., 2017.

Cotter, A., Keshet, J., and Srebro, N. Explicit approximations of the gaussian kernel. *CoRR*, abs/1109.4603, 2011.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Fino, B. J. and Algazi, V. R. Unified matrix treatment of the fast walsh-hadamard transform. *IEEE Transactions on Computers*, 25(11):1142–1146, 1976.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 457–468. Association for Computational Linguistics, 2016.

Gao, Y., Beijbom, O., Zhang, N., and Darrell, T. Compact bilinear pooling. *Proceedings of the 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 317–326, 2016.

Goldberg, Y. and Elhadad, M. splitsvm: Fast, space-efficient, non-heuristic, polynomial kernel computation for NLP applications. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pp. 237–240. The Association for Computer Linguistics, 2008.

Haagerup, U. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

Hamid, R., Xiao, Y., Gittens, A., and DeCoste, D. Compact random feature maps. In *Proceedings of the 31th International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 19–27. PMLR, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hogg, R., McKean, J., and Craig, A. *Introduction to Mathematical Statistics*, volume 8. Pearson, 2019.

Kar, P. and Karnick, H. Random feature maps for dot product kernels. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *JMLR Proceedings*, pp. 583–591. JMLR, 2012.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Larsen, K. G. and Nelson, J. Optimality of the johnson-lindenstrauss lemma. In *IEEE 58th Annual Symposium on Foundations of Computer Science*, pp. 633–638, 2017.

Latala, R. Estimation of moments of sums of independent real random variables. *The Annals of Probability*, 25(3): 1502–1513, 1997.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Meister, M., Sarlos, T., and Woodruff, D. Tight dimensionality reduction for sketching low degree polynomial kernels. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Milios, D., Camoriano, R., Michiardi, P., Rosasco, L., and Filippone, M. Dirichlet-based gaussian processes for

large-scale calibrated classification. In *Advances in Neural Information Processing Systems 31*, pp. 6008–6018. Curran Associates, Inc., 2018.

Park, K. I. *Fundamentals of Probability and Stochastic Processes with Applications to Communications*. Springer, 1st edition, 2018.

Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pp. 8026–8037. Curran Associates, Inc., 2019.

Pennington, J., Yu, F. X. X., and Kumar, S. Spherical random features for polynomial kernels. In *Advances in Neural Information Processing Systems 28*, pp. 1846–1854. Curran Associates, Inc., 2015.

Pham, N. and Pagh, R. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 239–247. Association for Computing Machinery, 2013.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pp. 1177–1184. Curran Associates Inc., 2007.

Rendle, S. Factorization machines. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, pp. 995–1000, 2010.

Russakovsky, O. et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Song, Z., Woodruff, D., Yu, Z., and Zhang, L. Fast sketching of polynomial kernels of polynomial degree. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9812–9823. PMLR, 2021.

Tropp, J. A. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(1-2):115–126, 2011.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, August 2012.

Wacker, J., Kanagawa, M., and Filippone, M. Improved random features for dot product kernels. *arXiv preprint arXiv:2201.08712*, 2022.

Weinberger, K. Q., Dasgupta, A., Langford, J., Smola, A. J., and Attenberg, J. Feature hashing for large scale multitask learning. In Danyluk, A. P., Bottou, L., and Littman, M. L. (eds.), *Proceedings of the 26th Annual International Conference on Machine Learning*, volume 382, pp. 1113–1120. ACM, 2009.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-ucsd birds 200. Technical report, Caltech, 2010. URL http://www.vision.caltech.edu/visipedia/CUB-200.html.

Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

STRUCTURE OF THE APPENDIX

- Appendix A contains proofs for the concentration results of Sections and 3.1 and 3.2 of the main paper.
- Appendix B contains variance derivations for non-structured CtR-sketches and proofs for the theorems in Section 3.3.
- Appendix C extends these variance derivations to (CtR-)ProductSRHT.
- Appendix D contains additional numerical experiments to complement Section 6 of the main paper.

A CONCENTRATION RESULTS

This section contains the proofs of Sections and 3.1 and 3.2 of the main paper. Many results in this section build on top of the work by Ahle et al. (2020). More precisely, we extend Ahle et al. (2020, Lem. 9, 11, 19, Thm. 42) to the case of CtR-sketches and show the improvements that our methods bring about. In particular, we derive absolute moments for CtR-sketches and show that these lead to sharper results.

A.1 Derivation of Moment Bounds and Proof of Lemma 3.1

We restate Lem. 3.1 here for ease of presentation. It is a complex extension of Ahle et al. (2020, Lem. 19).

Lemma A.1 (Absolute Moment Bound) *Let $t \geq 2, p \in \mathbb{N}$, $C_t > 0$, $\boldsymbol{x} \in \mathbb{R}^{d_1 \cdots d_p}$ and $\boldsymbol{w}_i \in \mathbb{C}^{d_i}$ for $i = 1, \dots, p$. If $\|\boldsymbol{w}_i^\top \boldsymbol{a}\|_{L^t} \leq C_t \|\boldsymbol{a}\|_2$ for all $\boldsymbol{a} \in \mathbb{R}^{d_i}$ and $\{\boldsymbol{w}_i\}_{i=1}^p$, then the following holds: $\|(\otimes_{i=1}^p \boldsymbol{w}_i)^\top \boldsymbol{x}\|_{L^t} \leq C_t^p \|\boldsymbol{x}\|_2$.*

Before proving Lem. A.1, we start by deriving the moment bounds C_t for (complex) Gaussian and Rademacher sketches.

A.1.1 Moment Bounds for Gaussian and Rademacher Sketches

W.l.o.g., we assume $\|\boldsymbol{a}\|_2 = 1$, since both sides of $\|\boldsymbol{w}_i^\top \boldsymbol{a}\|_{L^t} = C_t \|\boldsymbol{a}\|_2$ can be divided by $\|\boldsymbol{a}\|_2$.

Gaussian distribution For the simpler Gaussian case, we obtain C_t that is not only a tight upper bound, but an exact value for the t-th moment. That is, we have $\|\boldsymbol{w}_i^\top \boldsymbol{a}\|_{L^t} = C_t \|\boldsymbol{a}\|_2$. Moreover, we obtain values for $t > -1 \in \mathbb{R}$, and not only for even integers t.

We start with the real case, i.e., $\boldsymbol{w}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_i})$. Then $\boldsymbol{w}_i^\top \boldsymbol{a} \sim \mathcal{N}(0, 1)$, since Gaussians are closed under linear transformations. The t-th absolute moment of a Gaussian random variable is well-known. For $t > -1$, it is

$$\text{(Real case)} \qquad \mathbb{E}[|(\boldsymbol{w}_i)^\top \boldsymbol{a}|^t] = 2^{t/2} \Gamma\left(\frac{t+1}{2}\right) / \sqrt{\pi} \quad \iff \quad \|(\boldsymbol{w}_i)^\top \boldsymbol{a}\|_{L^t} = C_t = \sqrt{2} \pi^{-1/(2t)} \Gamma\left(\frac{t+1}{2}\right)^{1/t},$$

where $\Gamma(\cdot)$ is the Gamma function.

For the complex case, we have $\boldsymbol{w}_i \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_{d_i})$, which is equivalent to $\boldsymbol{w}_i = 1/\sqrt{2}(\boldsymbol{u}_i + \mathrm{i}\, \boldsymbol{v}_i)$ with $\boldsymbol{u}_i, \boldsymbol{v}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_i})$ being independent. Then we have

$$\mathbb{E}[|(\boldsymbol{w}_i)^\top \boldsymbol{a}|^t] = \mathbb{E}[|\sqrt{1/2}(\boldsymbol{u}_i^\top \boldsymbol{a} + \mathrm{i}\, \boldsymbol{v}_i^\top \boldsymbol{a})|^t] = (1/2)^{t/2}\, \mathbb{E}[(|\boldsymbol{u}_i^\top \boldsymbol{a}|^2 + |\boldsymbol{v}_i^\top \boldsymbol{a}|^2)^{t/2}]. \tag{8}$$

Now we observe that $\boldsymbol{u}_i^\top \boldsymbol{a}, \boldsymbol{v}_i^\top \boldsymbol{a} \sim \mathcal{N}(0, 1)$. So $|\boldsymbol{u}_i^\top \boldsymbol{a}|^2 + |\boldsymbol{v}_i^\top \boldsymbol{a}|^2$ is chi-square distributed with two degrees of freedom. The t'-th moment of a chi-square distributed variable X with $k \in \mathbb{N}$ degrees of freedom is (Hogg et al., 2019, Thm. 3.3.2.):

$$\mathbb{E}[X^{t'}] = 2^{t'} \Gamma(t' + k/2) / \Gamma(k/2), \quad \text{if} \quad t' > -k/2,$$

By setting $k = 2, t' = t/2$ and noting $\Gamma(1) = 1$, we obtain

$$\text{(Complex case)} \qquad \|(\boldsymbol{w}_i)^\top \boldsymbol{a}\|_{L^t} = C_t = \Gamma(t/2 + 1)^{1/t},$$

where $t > -2$ covers $t \geq 2$ required by the lemma.

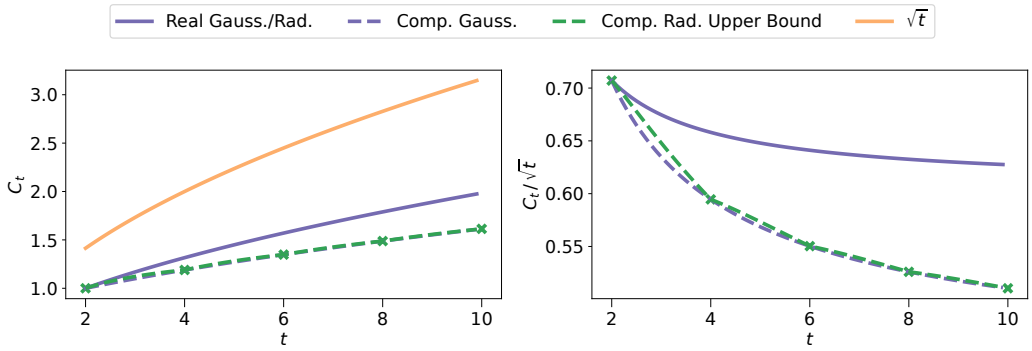


Figure 5: (Left) C_t values over $t \geq 2$. Values for the complex Rademacher case are interpolations between $t = 2k, k \in \mathbb{N}$. (Right) C_t values after division by $\sqrt{t}$.

Rademacher distribution For the real case, we can directly apply Khintchine's inequality stating $\|(\boldsymbol{w}_i)^\top \boldsymbol{a}\|_{L^t} \leq C_t \|\boldsymbol{a}\|_2$ with $0 < t < \infty$. Haagerup (1981) derived tight values for C_t yielding

$$\text{(Real case)} \quad \|(\boldsymbol{w}_i)^\top \boldsymbol{a}\|_{L^t} \leq C_t = \begin{cases} 1 & \text{for} \quad 0 < t \leq 2 \\ \sqrt{2}\pi^{-1/(2t)}\Gamma\left(\frac{t+1}{2}\right)^{1/t} & \text{for} \quad t > 2. \end{cases}$$

For the complex Rademacher case, we note that $|\exp(\mathrm{i}\pi/4)\boldsymbol{w}_i^\top \boldsymbol{a}| = |\boldsymbol{w}_i^\top \boldsymbol{a}|$ since $|\exp(\mathrm{i}\pi/4)| = 1$. The elements of $\boldsymbol{w}_i' := \exp(\mathrm{i}\pi/4)\boldsymbol{w}_i$ are then sampled from

$$\mathrm{Unif}(\{\exp(\mathrm{i}\pi/4), -\exp(\mathrm{i}\pi/4), \mathrm{i}\exp(\mathrm{i}\pi/4), -\mathrm{i}\exp(\mathrm{i}\pi/4)\}) = \frac{1}{\sqrt{2}}\mathrm{Unif}(\{1+\mathrm{i}, -1-\mathrm{i}, -1+\mathrm{i}, 1-\mathrm{i}\}).$$

We can thus rewrite $\boldsymbol{w}_i' = 1/\sqrt{2}(\boldsymbol{u}_i + \mathrm{i}\boldsymbol{v}_i)$ with $\boldsymbol{u}_i, \boldsymbol{v}_i$ having elements sampled i.i.d. from $\mathrm{Unif}(\{1, -1\})$. For $t = 2k, k \in \mathbb{N}$, we can further expand

$$\mathbb{E}[|\boldsymbol{w}_i^\top \boldsymbol{a}|^t] = (1/2)^{t/2}\mathbb{E}\left[(|\boldsymbol{u}_i^\top \boldsymbol{a}|^2 + |\boldsymbol{v}_i^\top \boldsymbol{a}|^2)^{t/2}\right] = (1/2)^{t/2}\sum_{n=0}^{t/2}\mathbb{E}\left[|\boldsymbol{u}_i^\top \boldsymbol{a}|^{2n}\right]\mathbb{E}\left[|\boldsymbol{v}_i^\top \boldsymbol{a}|^{2(t/2-n)}\right] \tag{9}$$

using the binomial theorem. Eq. 9 must be upper bounded by Eq. 8, since both have the same structure and, by Khintchine's inequality, the moments on the r.h.s. of Eq. 9 are upper bounded by the ones of the Gaussian distribution. Hence, we obtain

$$\text{(Complex case)} \qquad \|(\boldsymbol{w}_i)^\top \boldsymbol{a}\|_{L^t} \leq C_t = \Gamma(t/2+1)^{1/t} \quad \text{for} \quad t = 2k, k \in \mathbb{N}.$$

Since we have only derived C_t for $t = 2k, k \in \mathbb{N}$ for complex Rademacher sketches, we derive an interpolation strategy when $t \neq 2k$ in the following.

Interpolation of L^t-norms for complex Rademacher sketches For two random variables $X, Y \in \mathbb{C}$, Hölder's inequality gives

$$\|XY\|_{L^1} \leq \|X\|_{L^a}\|Y\|_{L^b} \quad \text{for} \quad a, b > 1 \quad \text{with} \quad 1/a + 1/b = 1.$$

We now define $a' := (b-a)/(b-t)$ and $b' := (b-a)/(t-a)$ for $a < t < b$, such that $a', b' > 1$ and $1/a' + 1/b' = 1$ are satisfied. We further have $a/a' + b/b' = t$. So we define a random variable $Z \in \mathbb{C}$ and obtain

$$\|Z^t\|_{L^1} = \|Z^{\frac{a}{a'}}Z^{\frac{b}{b'}}\|_{L^1} \leq \|Z^{\frac{a}{a'}}\|_{L^{a'}}\|Z^{\frac{b}{b'}}\|_{L^{b'}} = \|Z\|_{L^a}^{a/a'}\|Z\|_{L^b}^{b/b'} \iff \|Z\|_{L^t} \leq \|Z\|_{L^a}^{a/(ta')}\|Z\|_{L^b}^{b/(tb')}$$

via Hölder's inequality. We can thus set a and b equal to the closest even integer values below/above t and therefore obtain an upper bound for C_t. That is, we bound $\|Z\|_{L^a}^{a/(ta')} \leq C_a^{a/(ta')}$ and $\|Z\|_{L^b}^{b/(tb')} \leq C_b^{b/(tb')}$. Since $\|Z\|_{L^t} \leq C_t$ is assumed to be tight, we must have $\|Z\|_{L^t} \leq C_t \leq C_a^{a/(ta')}C_b^{b/(tb')}$.

The left plot of Fig. 5 shows C_t over t including our proposed interpolation for $t \neq 2k$ for the complex Rademacher case. We see that the upper bound matches the values for the complex Gaussian distribution almost exactly from $t \geq 4$. This is a strong indication for the fact that the actual C_t values are the same for both distributions, as we already showed for the real case. Furthermore, the upper bound for the complex Rademacher C_t values remains smaller than the C_t values for the real case. All functions grow more slowly than $\sqrt{t}$ as shown in the right plot of Fig. 5.

A.1.2 Proof of Lemma A.1 (Lemma 3.1 in the Main Paper)

Having derived the C_t values for (complex) Gaussian and Rademacher distributions, we are ready to prove Lem. A.1. The proof closely follows Ahle et al. (2020, Lem. 19), but extends it to the case of complex $\{\boldsymbol{w}_i\}_{i=1}^p$. We therefore provide the whole proof for completeness.

The proof is by induction. The initial case $p = 1$ is trivially fulfilled by the previous derivations. For the induction step, we assume that the claim is true for $p-1$. So we assume $\|(\otimes_{i=1}^{p-1} \boldsymbol{w}_i)^\top \boldsymbol{x}\|_{L^t} \leq C_t^{p-1} \|\boldsymbol{x}\|_2$. We now index the vector $\boldsymbol{x} \in \mathbb{R}^{d_1 \cdots d_p}$ in a tensorized fashion. So a single element of $\boldsymbol{x}$ is indexed as $x_{I_1,\dots,I_p}$ for indices $I_i \in \{1,\dots,d_i\}$. Let further $B_{I_1,\dots,I_{p-1}} = \sum_{I_p \in [d_p]} (\boldsymbol{w}_p)_{I_p} x_{I_1,\dots,I_p} \in \mathbb{C}$. Then $\mu_t := \mathbb{E}[|(\otimes_{i=1}^p \boldsymbol{w}_i)^\top \boldsymbol{x}|^t]$ yields

$$\mu_t = \mathbb{E}\left[\left|\sum_{I_1 \in [d_1],\dots,I_p \in [d_p]} \left(\prod_{i \in [p]} (\boldsymbol{w}_i)_{I_i}\right) x_{I_1,\dots,I_p}\right|^t\right] = \mathbb{E}\left[\left|\sum_{I_1 \in [d_1],\dots,I_{p-1} \in [d_{p-1}]} \left(\prod_{i \in [p-1]} (\boldsymbol{w}_i)_{I_i}\right) B_{I_1,\dots,I_{p-1}}\right|^t\right].$$

By the law of total expectation, this gives

$$\mu_t = \mathbb{E}\left[\mathbb{E}\left[\left|\sum_{I_1 \in [d_1],\dots,I_{p-1} \in [d_{p-1}]} \left(\prod_{i \in [p-1]} (\boldsymbol{w}_i)_{I_i}\right) B_{I_1,\dots,I_{p-1}}\right|^t \middle| \boldsymbol{w}_p\right]\right].$$

By the induction assumption, we get

$$\mu_t \leq C_t^{t(p-1)} \mathbb{E}\left[\left|\sum_{I_1 \in [d_1],\dots,I_{p-1} \in [d_{p-1}]} |B_{I_1,\dots,I_{p-1}}|^2\right|^{1/2 \cdot t}\right] = C_t^{t(p-1)} \left\|\sum_{I_1 \in [d_1],\dots,I_{p-1} \in [d_{p-1}]} |B_{I_1,\dots,I_{p-1}}|^2\right\|_{L^{t/2}}^{t/2}.$$

Since $t/2 \geq 1$, we use Minkowski's inequality (triangle inequality for the L_t-norm) to move the norm inside the sum:

$$\mu_t \leq C_t^{t(p-1)} \left(\sum_{I_1 \in [d_1],\dots,I_{p-1} \in [d_{p-1}]} \||B_{I_1,\dots,I_{p-1}}|^2\|_{L^{t/2}}\right)^{t/2} = C_t^{t(p-1)} \left(\sum_{I_1 \in [d_1],\dots,I_{p-1} \in [d_{p-1}]} \|B_{I_1,\dots,I_{p-1}}\|_{L^t}^2\right)^{t/2}.$$

Now, recall that $B_{I_1,\dots,I_{p-1}}$ is a weighted sum with weights $\{(\boldsymbol{w}_p)_{I_p}\}_{I_p=1}^{d_p}$. So we can use the initial assumption $\|\boldsymbol{w}_p^\top \boldsymbol{a}\|_{L^t} \leq C_t \|\boldsymbol{a}\|_2$ for all $\boldsymbol{a} \in \mathbb{R}^{d_p}$. Therefore, we have $\|B_{I_1,\dots,I_{p-1}}\|_{L^t} \leq C_t (\sum_{I_p \in [d_p]} (x_{I_1,\dots,I_p})^2)^{1/2}$, and finally

$$\mu_t \leq C_t^{t(p-1)} \left(\sum_{I_1 \in [d_1],\dots,I_{p-1} \in [d_{p-1}]} C_t^2 \sum_{I_p \in [d_p]} (x_{I_1,\dots,I_p})^2\right)^{t/2} = C_t^{tp} \|\boldsymbol{x}\|_2^{2 \cdot t/2} = C_t^{tp} \|\boldsymbol{x}\|_2^t,$$

which proves the claim $\|(\otimes_{i=1}^p \boldsymbol{w}_i)^\top \boldsymbol{x}\|_{L^t} \leq C_t^p \|\boldsymbol{x}\|_2$. ∎

When $\boldsymbol{x} = \otimes_{i=1}^p \boldsymbol{x}_i$ is a tensor product for some $\boldsymbol{x}_i \in \mathbb{R}^{d_i}$, $i = 1,\dots,p$. Then we have

$$\mathbb{E}[|(\otimes_{i=1}^p \boldsymbol{w}_i)^\top (\otimes_{i=1}^p \boldsymbol{x}_i)|^t] = \mathbb{E}\left[\left|\prod_{i=1}^p \boldsymbol{w}_i^\top \boldsymbol{x}_i\right|^t\right] = \prod_{i=1}^p \mathbb{E}\left[\left|\boldsymbol{w}_i^\top \boldsymbol{x}_i\right|^t\right] \leq \prod_{i=1}^p C_t^t \|\boldsymbol{x}_i\|_2^t = C_t^{tp} \|\otimes_{i=1}^p \boldsymbol{x}_i\|_2^t.$$

Since $\mathbb{E}[|\boldsymbol{w}_i^\top \boldsymbol{x}_i|^t] \leq C_t^t \|\boldsymbol{x}_i\|_2^t$ is tight by the assumption that C_t are tight constants, the bound in Lem. A.1 becomes tight too in this case.

A.2 Proof of Theorem 3.2

We prove Thm. 3.2 for the more general case of $\delta \in (0, \exp(-2p\gamma))$, for which we introduce an additional variable $\gamma > 0$:

Theorem A.2 *Let $\epsilon, \gamma > 0, p \in \mathbb{N}, \delta \in (0, \exp(-2p\gamma)), \boldsymbol{x} \in \mathbb{R}^{d_1 \cdots d_p}, \boldsymbol{S} = (\boldsymbol{s}_1, \dots, \boldsymbol{s}_D)^\top \in \mathbb{C}^{D \times d_1 \cdots d_p}$ with $\boldsymbol{s}_\ell = \otimes_{i=1}^p \boldsymbol{w}_{i,\ell}/\sqrt{D}$ and $\boldsymbol{w}_{i,\ell} \in \mathbb{C}^{d_i}$ be i.i.d. Gaussian/Rademacher samples as in Lem. 3.1. In order to guarantee*

$$\Pr\left\{ \left| \|\boldsymbol{S}\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2 \right| \leq \epsilon \|\boldsymbol{x}\|_2^2 \right\} \geq 1 - \delta, \quad \textit{we need}$$

$$D = \mathcal{O}\left(\max\left\{ (C_4 e^{\gamma/2})^{4p} \left(\frac{\log(1/\delta)}{p\gamma} \right) \epsilon^{-2}, \quad (C_4^2 e / 2e^{\gamma})^p \left(\frac{\log(1/\delta)}{p\gamma} \right)^p \epsilon^{-1} \right\} \right),$$

where C_4 equals $3^{1/4}(2^{1/4})$ for the real (complex) Gaussian/Rademacher distribution, as derived in Lem. 3.1.

Setting $\gamma = 1/p$ yields the formulation of Thm. 3.2. Alternatively, we may allow $\delta \in (0, 1)$ by letting $\gamma \propto 1/p$ go towards zero. However, this leads to a worse dependence on p, since $(1/\gamma)^p$ becomes arbitrarily large.

Proof The proof is an extension of Ahle et al. (2020, Appendix A.2) to the case of complex $\boldsymbol{S}$. It thus makes use of Lem. 3.1 in order to prove the theorem. Moreover, we make use of interpolated values for C_t when $t \neq 2k$ for any $k \in \mathbb{N}$ by using an upper bound on C_t in this case, as shown in Section A.1. Crucially however, this upper-bound interpolation does not harm the sharpness of our results. Moreover, the original proof by Ahle et al. (2020, Appendix A.2) requires $t = \frac{\log(1/\delta)}{p\gamma} \geq 4$, which we relax to $t = \frac{\log(1/\delta)}{p\gamma} > 2$ to allow for a larger range of error probabilities, i.e., we require $\delta \in (0, \exp(-2p\gamma))$ instead of $\delta \in (0, \exp(-4p\gamma))$ in the theorem. We provide the entire modified proof here for completeness.

Our goal is to show that $\| \|\boldsymbol{S}\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2 \|_{L^t} \leq \delta^{1/t} \epsilon \|\boldsymbol{x}\|_2^2$ holds. Then we can apply Markov's inequality: $\Pr\{X \geq a\} \leq \mathbb{E}[X]/a$ for $a > 0$, where we set $X = | \|\boldsymbol{S}\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2 |^t$ and $a = \epsilon^t \|\boldsymbol{x}\|_2^{2t}$ to obtain

$$\Pr\{ | \|\boldsymbol{S}\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2 |^t \geq \epsilon^t \|\boldsymbol{x}\|_2^{2t} \} \leq \delta \quad \iff \quad \Pr\{ | \|\boldsymbol{S}\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2 | \leq \epsilon \|\boldsymbol{x}\|_2^2 \} \geq 1 - \delta. \tag{10}$$

Without loss of generality, we can assume $\|\boldsymbol{x}\|_2 = 1$ from now onward, since

$$\left\| \|\boldsymbol{S}\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2 \right\|_{L^t} \leq \epsilon \delta^{1/t} \|\boldsymbol{x}\|_2^2 \iff \left\| \left\| \boldsymbol{S}\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} \right) \right\|_2^2 - 1 \right\|_{L^t} \leq \epsilon \delta^{1/t}.$$

In order to prove $\| \|\boldsymbol{S}\boldsymbol{x}\|_2^2 - 1 \|_{L^t} \leq \epsilon \delta^{1/t}$, we write $\boldsymbol{S} = (\boldsymbol{s}_1, \dots, \boldsymbol{s}_D)^\top$ with $\boldsymbol{s}_\ell = (\otimes_{i=1}^p \boldsymbol{w}_{i,\ell})/\sqrt{D}$ and $\boldsymbol{w}_{i,\ell} \in \mathbb{C}^{d_i}$ i.i.d. as in Section 3.1. So we can reformulate

$$\left\| \|\boldsymbol{S}\boldsymbol{x}\|_2^2 - 1 \right\|_{L^t} = \left\| \left(\frac{1}{D} \sum_{\ell=1}^D \left| (\otimes_{i=1}^p \boldsymbol{w}_{i,\ell})^\top \boldsymbol{x} \right|^2 \right) - 1 \right\|_{L^t} = \left\| \frac{1}{D} \sum_{\ell=1}^D Z_\ell \right\|_{L^t} \tag{11}$$

with $Z_\ell := |(\otimes_{i=1}^p \boldsymbol{w}_{i,\ell})^\top \boldsymbol{x}|^2 - 1$ being i.i.d. random variables with zero mean, since $\mathbb{E}[|(\otimes_{i=1}^p \boldsymbol{w}_{i,\ell})^\top \boldsymbol{x}|^2] = \|\boldsymbol{x}\|_2^2 = 1$. Next, we bound $\|Z_\ell\|_{L^t}$ using Minkowski's inequality:

$$\|Z_\ell\|_{L^t} = \| |(\otimes_{i=1}^p \boldsymbol{w}_{i,\ell})^\top \boldsymbol{x}|^2 - 1 \|_{L^t} \leq \| |(\otimes_{i=1}^p \boldsymbol{w}_{i,\ell})^\top \boldsymbol{x}|^2 \|_{L^t} + \| -1 \|_{L^t} = \|(\otimes_{i=1}^p \boldsymbol{w}_{i,\ell})^\top \boldsymbol{x}\|_{L^{2t}}^2 + 1. \tag{12}$$

We can further bound $\|(\otimes_{i=1}^p \boldsymbol{w}_{i,\ell})^\top \boldsymbol{x}\|_{L^{2t}} \leq C_{2t}^p$ for any $t \geq 1$ by Lem. 3.1. Precise values $C_{t'}$ are derived in the Lemma, except for the complex Rademacher case for which we provide values for $t' = 2k, k \in \mathbb{N}$. For $t' \neq 2k$, we can use the upper-bound interpolation

$$C_{t'} \leq C_a^{\frac{a(b-t')}{t'(b-a)}} C_b^{\frac{b(t'-a)}{t'(b-a)}} \quad \text{with} \quad a < t' < b,$$

where we choose a and b to be the closest even integer values below and above t, respectively.

As shown in Fig. 5, $C_{t'}$ grows more slowly than $\sqrt{t}$. $f(t') := C_{t'}/\sqrt{t'}$ is thus a monotonically decreasing function. Now we set $t' = 2t > 4$ by our initial assumption $t > 2$, and we obtain $f(t') \leq C_4/\sqrt{4}$. We can thus bound $C_{2t} \leq C_4\sqrt{2t}/\sqrt{4}$. This eventually allows us to bound $\|Z_\ell\|_{L^t} \leq (C_4^2 t/2)^p + 1$ for all $\ell \in \{1, \dots, D\}$ in Eq. 12. Notably, the fact of having interpolated $C_{t'}$ by using an upper bound for complex Rademacher sketches has not influenced this result, since $f(t') \leq C_4\sqrt{4}$ would remain valid even if the $C_{t'}$ values were smaller for $t' \neq 2k$.

In order to bound $\| \frac{1}{D} \sum_{\ell=1}^D Z_\ell \|_{L^t}$ (11), we need Latala's inequality:

Lemma A.3 (Latala (1997), Corollary 2) *If $p \geq 2$ and $X, X_1, \dots, X_n$ are i.i.d. symmetric random variables, then we have*

$$\|X_1 + \cdots + X_n\|_{L^t} \sim \sup\left\{ \frac{t}{s}\left(\frac{n}{t}\right)^{1/s} \|X\|_{L^s} \,\middle|\, \max\{2, t/n\} \leq s \leq t \right\}.$$

Here, $f(x) \sim g(x)$ means $c_1 g(x) \leq f(x) \leq c_2 g(x)$ for all x and some universal constants c_1, c_2. By Latala (1997, Remark 2), the lemma is also valid for zero-mean random variables with slightly worse constants $1/2\, c_1$ and $2\, c_2$.

Recall that $\|Z_\ell\|_{L^t} \leq (C_4^2 t/2)^p + 1 \leq c_3 (C_4^2 t/2)^p$ for some $c_3 > 0$. W.l.o.g., we can set $c_3 = 1$ when substituting $\|X\|_{L^s}$ by $c_3 (C_4^2 t/2)^p$ inside Lem. A.3. The functional form $h(t) := (Kt)^p$ with $K := C_4^2/2 > 0$ allows us to greatly simplify Lem. A.3 using the following corollary.

Corollary A.4 (Ahle et al. (2020), Corollary 38) *If $\|X\|_{L^s} \sim (Ks)^p$ for some $p \geq 1, K > 0$, then the supremum in Lem. A.3 is attained for the minimal and maximal case of s, i.e., $s = \max\{2, t/n\}$ and $s = t$. Lem. A.3 then becomes*

$$\|X_1 + \cdots + X_n\|_{L^t} \sim K^p \max\left\{ \sqrt{tn}\, 2^p, \left(\frac{n}{t}\right)^{1/t} t^p \right\}.$$

Using Cor. A.4 and setting $K = C_4^2/2$, we obtain the following bound on $\|\frac{1}{D}\sum_{\ell=1}^D Z_\ell\|_{L^t}$ (11):

$$\left\| \frac{1}{D}\sum_{\ell=1}^D Z_\ell \right\|_{L^t} \sim \frac{1}{D}(C_4^2/2)^p \max\left\{ \sqrt{tD}\, 2^p, \left(\frac{D}{t}\right)^{1/t} t^p \right\} = \max\{\underbrace{C_4^{2p}\sqrt{t/D}}_{(1)}, \underbrace{(C_4^2/2\,t)^p (D/t)^{1/t}/D}_{(2)}\}. \tag{13}$$

Recall that our goal is to provide a condition on D for which $\|\frac{1}{D}\sum_{\ell=1}^D Z_\ell\|_{L^t} \leq \epsilon\delta^{1/t}$ holds. Since we can freely choose $t > 2$, we set it to $t = \frac{\log(1/\delta)}{p\gamma} > 2$ for some $\gamma > 0$ from now onward. Then, it is only left to show that terms (1) and (2) in Eq. 13 are upper-bounded by $\epsilon\delta^{1/t}$. We start with the simpler case (1).

Analysis of case (1) Setting $D \geq (C_4 e^{\gamma/2})^{4p} \frac{\log(1/\delta)}{p\gamma} \epsilon^{-2}$ directly gives

$$(1) = (C_4)^{2p}\sqrt{\frac{\log(1/\delta)}{p\gamma D}} \leq (C_4)^{2p}\sqrt{\frac{\log(1/\delta)\epsilon^2}{p\gamma (C_4 e^{\gamma/2})^{4p}\frac{\log(1/\delta)}{p\gamma}}} = \epsilon e^{-\gamma p} = \epsilon e^{-\gamma p t/t} = \epsilon\delta^{1/t}.$$

Analysis of case (2) For a simpler analysis, we start by upper bounding $D^{1/t}$ in case (2). For this purpose, we study the condition in which (2) $\geq$ (1) s.t. our error is upper bounded by (2). We have

$$\underbrace{(C_4^2)^p\sqrt{t/D}}_{(1)} \leq \underbrace{(C_4^2/2\,t)^p (D/t)^{1/t}/D}_{(2)} \leq (C_4^2/2\,t)^p D^{1/t}/D$$

$$\iff 2^p t^{1/2-p} \leq D^{1/t-1/2} \iff D^{1/t} \leq \left(\frac{t}{2}\right)^{\frac{2p-1}{t-2}} \left(\frac{1}{2}\right)^{\frac{1}{t-2}} \overset{(t>2)}{\leq} \left(\frac{t}{2}\right)^{\frac{2p-1}{t-2}} \overset{(t\to 2)}{\leq} \exp(p - 1/2).$$

Thus, if (2) $\geq$ (1), we have (2) $\leq e^{-1/2}(C_4^2/2\,et)^p/D$. Setting $D \geq (C_4^2/2\,tee^\gamma)^p \epsilon^{-1} e^{-1/2}$ finally yields

$$(2) \leq e^{-1/2}(C_4^2/2\,et)^p/D \leq \frac{e^{-1/2}(C_4^2/2\,et)^p}{(C_4^2/2\,tee^\gamma)^p \epsilon^{-1} e^{-1/2}} = \epsilon e^{-\gamma p} = \epsilon\delta^{1/t}.$$

Setting D to the maximum value of the conditions of case (1) and (2) ensures that $\|\frac{1}{D}\sum_{\ell=1}^D Z_\ell\|_{L^t} \leq \epsilon\delta^{1/t}$ in Eq. 13. ■

A.3 A Comparison with Wacker et al. (2022), Theorem 3.4

Wacker et al. (2022, Theorem 3.4) provide an error bound relative to the L1-norm of the form

$$\Pr\left\{ |\|\boldsymbol{S}\boldsymbol{a}^{\otimes p}\|_2^2 - \|\boldsymbol{a}^{\otimes p}\|_2^2| \leq \epsilon \|\boldsymbol{a}\|_1^{2p} \right\} \geq 1 - \delta,$$

where $\boldsymbol{S} \in \mathbb{C}^{D \times d^p}$ is a complex Rademacher sketch and $\boldsymbol{a} \in \mathbb{R}^d$.

Bounding the error $\epsilon > 0$ relative to the L1-norm of $\boldsymbol{a}$ instead of the L2-norm is problematic. To see this, consider the vector $\boldsymbol{a} = (1, \dots, 1)^\top / \sqrt{d} \in \mathbb{R}^d$. It has $\|\boldsymbol{a}\|_2 = 1$ and $\|\boldsymbol{a}\|_1 = \sqrt{d}$. In this case, we have $\|\boldsymbol{a}\|_1^{2p} = d^p \|\boldsymbol{a}\|_2^{2p}$. Since the bound by Wacker et al. (2022) requires $D = \mathcal{O}(\epsilon^{-2})$, this would translate into a guarantee of $D = \mathcal{O}(d^{2p})$ to bound the error relative to $\|\boldsymbol{a}\|_2^{2p}$. Hence, D is already larger than the dimension d^p of $\boldsymbol{a}^{\otimes p}$, which defeats the purpose of dimensionality reduction.

A.4 Proof of Corollary 3.3 (Approximate Matrix Product)

We want to show that

$$\Pr\left\{|(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{x})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{y}) - \boldsymbol{x}^\top\boldsymbol{y}| \le \epsilon \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2\right\} \ge 1 - \delta$$

holds for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{d_1 \cdots d_p}$ and $\boldsymbol{S}_{\mathrm{CtR}} := (\mathrm{Re}\{\boldsymbol{s}_1\}, \dots, \mathrm{Re}\{\boldsymbol{s}_D\}, \mathrm{Im}\{\boldsymbol{s}_1\}, \dots, \mathrm{Im}\{\boldsymbol{s}_D\})^\top \in \mathbb{R}^{2D \times d_1 \cdots d_p}$ with $\boldsymbol{S} = (\boldsymbol{s}_1, \dots, \boldsymbol{s}_D)^\top$ being the same as in Thm. 3.2.

Proof Our first goal is to show that the following holds:

$$\|(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{x})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{y}) - \boldsymbol{x}^\top\boldsymbol{y}\|_{L^t} \le \epsilon\delta^{1/t} \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 .$$

W.l.o.g., we can assume $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 = 1$ from now onward, since both sides of the inequality can be divided by $\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2$.

In Section A.2, we have shown that $\| \|\boldsymbol{S}\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2 \|_{L^t} \le \delta^{1/t}\epsilon \|\boldsymbol{x}\|_2^2$ holds for any $\boldsymbol{x} \in \mathbb{R}^{d_1 \cdots d_p}$ and $t > 2$. Recall that $\|\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{x}\|_2^2 = \sum_{\ell=1}^D \mathrm{Re}\{\boldsymbol{s}_\ell^\top \boldsymbol{x}\}^2 + \mathrm{Im}\{\boldsymbol{s}_\ell^\top \boldsymbol{x}\}^2 = \|\boldsymbol{S}\boldsymbol{x}\|_2^2$, which already implies $\| \|\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2 \|_{L^t} \le \delta^{1/t}\epsilon \|\boldsymbol{x}\|_2^2$.

The rest of the proof follows Ahle et al. (2020, Lem. 9). For two vectors $\boldsymbol{a}, \boldsymbol{b}$, we have $\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 = \|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2 - 2(\boldsymbol{a}^\top\boldsymbol{b})$ and $\|\boldsymbol{a} + \boldsymbol{b}\|_2^2 = \|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2 + 2(\boldsymbol{a}^\top\boldsymbol{b})$. Being combined, this gives $\boldsymbol{a}^\top\boldsymbol{b} = \left(\|\boldsymbol{a} + \boldsymbol{b}\|_2^2 - \|\boldsymbol{a} - \boldsymbol{b}\|_2^2\right)/4$. Hence

$$\begin{aligned}
\left\|(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{x})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{y}) - \boldsymbol{x}^\top\boldsymbol{y}\right\|_{L^t} &= \left\|\|\boldsymbol{S}_{\mathrm{CtR}}(\boldsymbol{x} + \boldsymbol{y})\|_2^2 - \|\boldsymbol{S}_{\mathrm{CtR}}(\boldsymbol{x} - \boldsymbol{y})\|_2^2 - \|\boldsymbol{x} + \boldsymbol{y}\|_2^2 + \|\boldsymbol{x} - \boldsymbol{y}\|_2^2\right\|_{L^t}/4 \\
&\le \left(\left\|\|\boldsymbol{S}_{\mathrm{CtR}}(\boldsymbol{x} + \boldsymbol{y})\|_2^2 - \|\boldsymbol{x} + \boldsymbol{y}\|_2^2\right\|_{L^t} + \left\|\|\boldsymbol{S}_{\mathrm{CtR}}(\boldsymbol{x} - \boldsymbol{y})\|_2^2 - \|\boldsymbol{x} - \boldsymbol{y}\|_2^2\right\|_{L^t}\right)/4 \\
&\le \epsilon\delta^{1/t}\left(\|\boldsymbol{x} + \boldsymbol{y}\|_2^2 + \|\boldsymbol{x} - \boldsymbol{y}\|_2^2\right)/4 = \epsilon\delta^{1/t}\left(\|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2\right)/2 = \epsilon\delta^{1/t}.
\end{aligned}$$

To conclude the proof, we apply Markov's inequality $\Pr(X \ge a) \le \mathbb{E}[X]/a$ with $a > 0$, as we did in Section A.2.

In the cases of matrices $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \in \mathbb{R}^{d_1 \cdots d_p \times n}, \boldsymbol{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_m) \in \mathbb{R}^{d_1 \cdots d_p \times m}$, we set $a = \epsilon^2\|\boldsymbol{X}\|_F^2\|\boldsymbol{Y}\|_F^2$ and $X = \|(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{X})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{Y}) - \boldsymbol{X}^\top\boldsymbol{Y}\|_F^2$ inside the inequality. Then we get

$$\Pr\left\{(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{X})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{Y}) - \boldsymbol{X}^\top\boldsymbol{Y}\|_F^2 \ge \epsilon^2 \|\boldsymbol{X}\|_F^2 \|\boldsymbol{Y}\|_F^2\right\} \le \frac{\epsilon^2\delta \sum_{i=1}^n \sum_{j=1}^m \|\boldsymbol{x}_i\|_2^2 \|\boldsymbol{y}_j\|_2^2}{\epsilon^2\|\boldsymbol{X}\|_F^2\|\boldsymbol{Y}\|_F^2} = \delta \tag{14}$$

when $\boldsymbol{S}_{\mathrm{CtR}}$ has $2D$ rows with D being the same as in Thm. 3.2. ■

A.5 From Approximate Matrix Products to Subspace Embeddings

We now use the inequality (14) derived in Section A.4 to bound the spectral approximation error of the polynomial kernel matrix. We define the target gram matrix $\boldsymbol{K} := (\boldsymbol{X}^{\otimes p})^\top\boldsymbol{X}^{\otimes p} + \lambda\boldsymbol{I}_n$, where $\boldsymbol{X}^{\otimes p} = (\boldsymbol{x}_1^{\otimes p}, \dots, \boldsymbol{x}_n^{\otimes p}) \in \mathbb{R}^{d^p \times n}$ is a matrix containing the polynomial feature maps of the data points $\{\boldsymbol{x}_i\}_{i=1}^n$, and $\lambda \ge 0$ is a regularization parameter. Our task is to determine D for which we can guarantee

$$(1 - \epsilon)(\boldsymbol{K} + \lambda\boldsymbol{I}_n) \preceq (\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{X}^{\otimes p})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{X}^{\otimes p}) + \lambda\boldsymbol{I}_n \preceq (1 + \epsilon)(\boldsymbol{K} + \lambda\boldsymbol{I}_n) \tag{15}$$

with probability at least $1 - \delta$.

Proof We rephrase Ahle et al. (2020, Lem. 11) here for the case $\lambda > 0$. This ensures that $\boldsymbol{K} + \lambda\boldsymbol{I}_n$ is positive definite and $(\boldsymbol{K} + \lambda\boldsymbol{I}_n)^{-1/2}$ exists. The same result for $\lambda = 0$ can then be obtained using Fatou's as shown in the original lemma.

By Tropp (2012, Prop. 2.1.1.), left and right multiplying the spectral inequality (15) by $(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1/2}$ does not change the positive semi-definite order. So (15) becomes

$$(1-\epsilon)\boldsymbol{I}_n \preceq (\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{X}^{\otimes p}(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1/2})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{X}^{\otimes p}(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1/2})^\top + \lambda(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1} \preceq (1+\epsilon)\boldsymbol{I}_n,$$

which is equivalent to

$$\|(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{X}^{\otimes p}(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1/2})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{X}^{\otimes p}(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1/2})^\top + \lambda(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1} - \boldsymbol{I}_n\|_2 \leq \epsilon. \tag{16}$$

Now we define $\boldsymbol{Z} := \boldsymbol{X}^{\otimes p}((\boldsymbol{X}^{\otimes p})^\top\boldsymbol{X}^{\otimes p}+\lambda\boldsymbol{I}_n)^{-1/2}$ so that

$$\begin{aligned}\boldsymbol{Z}^\top\boldsymbol{Z} &= (\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1/2}\boldsymbol{K}(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1/2}\\ &= (\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1/2}(\boldsymbol{K}+\lambda\boldsymbol{I}_n-\lambda\boldsymbol{I}_n)(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1/2}\\ &= \boldsymbol{I}_n - \lambda(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1}.\end{aligned}$$

Then (16) becomes $\|(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{Z})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{Z}) - \boldsymbol{Z}^\top\boldsymbol{Z}\|_2 \leq \epsilon$. and we can apply our bound on the Frobenius norm error (14), since it holds for *any* $\boldsymbol{X}, \boldsymbol{Y}$. So we can set $\boldsymbol{X}=\boldsymbol{Y}=\boldsymbol{Z}\in\mathbb{R}^{d^p\times n}$ and get:

$$\begin{aligned}&\Pr\{\|(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{Z})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{Z}) - \boldsymbol{Z}^\top\boldsymbol{Z}\|_2 \geq \epsilon\|\boldsymbol{Z}\|_F\|\boldsymbol{Z}\|_F\}\\ &\leq \Pr\{\|(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{Z})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{Z}) - \boldsymbol{Z}^\top\boldsymbol{Z}\|_F \geq \epsilon\|\boldsymbol{Z}\|_F\|\boldsymbol{Z}\|_F\} \leq \delta.\end{aligned} \tag{17}$$

Now we have that

$$\|\boldsymbol{Z}\|_F^2 = \mathrm{tr}(\boldsymbol{Z}^\top\boldsymbol{Z}) = \mathrm{tr}(\boldsymbol{I}_n - \lambda(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1})$$

and $\mathrm{tr}(\boldsymbol{Z}^\top\boldsymbol{Z}) = \sum_{i=1}^n \lambda_i(\boldsymbol{Z}^\top\boldsymbol{Z})$ is the sum over eigenvalues $\lambda_i(\boldsymbol{Z}^\top\boldsymbol{Z})$. This gives

$$\mathrm{tr}(\boldsymbol{Z}^\top\boldsymbol{Z}) = \sum_{i=1}^n 1 - \frac{\lambda}{\lambda+\lambda_i(\boldsymbol{K})} = \sum_{i=1}^n \frac{\lambda_i(\boldsymbol{K})}{\lambda_i(\boldsymbol{K})+\lambda} = \mathrm{tr}(\boldsymbol{K}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}) =: s_\lambda(\boldsymbol{K}),$$

where $0 \leq s_\lambda(\boldsymbol{K}) \leq n$ is the λ-statistical dimension of $\boldsymbol{K}$.

Substituting $\epsilon = \epsilon'/\|\boldsymbol{Z}\|_F^2 = \epsilon' s_\lambda(\boldsymbol{K})^{-1}$ for some $\epsilon' > 0$ in (17) ensures that $\|(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{Z})^\top(\boldsymbol{S}_{\mathrm{CtR}}\boldsymbol{Z}) - \boldsymbol{Z}^\top\boldsymbol{Z}\|_2 \leq \epsilon'$ is satisfied with probability at least $1-\delta$, when $\boldsymbol{S}_{\mathrm{CtR}}$ has $2Ds_\lambda(\boldsymbol{K})^2$ rows, where D is the same as in Thm. 3.2. ■

B VARIANCE OF COMPLEX-TO-REAL SKETCHES

In this section, we derive the variances of non-structured CtR-sketches.

B.1 The structure of CtR variances

We start by deriving the general variance structure of CtR-sketches that we will frequently refer to later on. For a complex random variable $z = a + \mathrm{i}\, b$ with $a, b \in \mathbb{R}$, we have $|z|^2 = a^2 + b^2$ and $\mathrm{Re}\{z^2\} = a^2 - b^2$. Combining both equations gives $a^2 = \frac{1}{2}(|z|^2 + \mathrm{Re}\{z^2\})$. The scalar a is real-valued and its variance $\mathbb{V}[a] = \mathbb{E}[a^2] - \mathbb{E}[a]^2$ is thus

$$\mathbb{V}[a] = \frac{1}{2}\mathrm{Re}\{\mathbb{E}[|z|^2] + \mathbb{E}[z^2] - 2\mathbb{E}[a]^2\}. \tag{18}$$

Let $\Phi_{\mathrm{C}} : \mathbb{R}^d \to \mathbb{C}^D$ be a complex polynomial sketch as defined in Eq. 4 and be $\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y}) = \Phi_{\mathrm{C}}(\boldsymbol{x})^\top \overline{\Phi_{\mathrm{C}}(\boldsymbol{y})} \in \mathbb{C}$ the associated approximate kernel for some $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. As the kernel estimate is an unbiased estimate of the real-valued target kernel $k(\boldsymbol{x}, \boldsymbol{y})$, we have

$$\mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] = \mathbb{E}[\mathrm{Re}\{\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})\}] + \mathrm{i} \cdot \mathbb{E}[\mathrm{Im}\{\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})\}] = k(\boldsymbol{x}, \boldsymbol{y}).$$

From this it follows that $\mathbb{E}[\mathrm{Im}\{\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})\}] = 0$ and therefore $\mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] = \mathbb{E}[\mathrm{Re}\{\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})\}] = k(\boldsymbol{x}, \boldsymbol{y})$. Setting $z = \hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})$ and $a = \mathrm{Re}\{\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})\} =: \hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})$ in Eq. 18 yields

$$\begin{aligned}
\mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})] &= \frac{1}{2}\mathrm{Re}\{\mathbb{E}[|\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})|^2] + \mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})^2] - 2\mathbb{E}[\mathrm{Re}\{\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})\}]^2\} \\
&= \frac{1}{2}\mathrm{Re}\{\mathbb{E}[|\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})|^2] + \mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})^2] - 2\mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]^2\} \\
&= \frac{1}{2}\mathrm{Re}\{\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] + \mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]\},
\end{aligned}$$

where $\mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] := \mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})^2] - \mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] \in \mathbb{C}$ is called the *pseudo-variance* of $\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})$ (Park, 2018, Chapter 5). In fact, we show next that $\mathrm{Im}\{\mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]\} = 0$ for all the sketches discussed in this work. Hence, we can also write $\mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})] = \frac{1}{2}(\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] + \mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})])$ for them since $\mathbb{V}[z] \in \mathbb{R}$ for any $z \in \mathbb{C}$. In order to determine $\mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})]$, we thus work out $\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]$ and $\mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]$ for Gaussian, Rademacher and ProductSRHT sketches in the following.

B.2 Gaussian and Rademacher sketches

In this section, we work out the variance of Gaussian and Rademacher CtR-sketches. For a set of D i.i.d. random feature samples, we have

$$\mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})] = \mathbb{V}[\mathrm{Re}\{\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})\}] = \mathbb{V}\left[\mathrm{Re}\{\Phi_{\mathrm{C}}(\boldsymbol{x})^\top \overline{\Phi_{\mathrm{C}}(\boldsymbol{y})}\}\right] = \frac{1}{D^2}\sum_{\ell=1}^{D} \mathbb{V}\left[\mathrm{Re}\left\{\prod_{i=1}^{p} (\boldsymbol{w}_{i,\ell}^\top \boldsymbol{x})\overline{(\boldsymbol{w}_{i,\ell}^\top \boldsymbol{y})}\right\}\right]. \tag{19}$$

As $\{\boldsymbol{w}_{i,\ell}\}_{\ell=1}^{D}$ are i.i.d., the variance terms are equal for each ℓ in Eq. 19 and $\mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})] \propto 1/D$. We can therefore assume $D = 1$ and drop the index ℓ for simplicity in the following. We then rescale the variances by $1/D$ later.

As our estimator is unbiased, we have $\mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] = k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^\top \boldsymbol{y})^p$. Thus, we only need to work out $\mathbb{E}[|\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})|^2]$ and $\mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})^2]$ for the variance and pseudo-variance, respectively.

Pseudo-Variance We start with $\mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})^2]$ to derive the pseudo-variance $\mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]$ after.

$$\mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})^2] = \mathbb{E}\left[\left(\prod_{i=1}^{p} \boldsymbol{w}_i^\top \boldsymbol{x}\overline{\boldsymbol{w}_i^\top \boldsymbol{y}}\right)^2\right] = \prod_{i=1}^{p} \mathbb{E}\left[(\boldsymbol{w}_i^\top \boldsymbol{x})^2(\overline{\boldsymbol{w}_i^\top \boldsymbol{y}})^2\right] = \mathbb{E}\left[(\boldsymbol{w}^\top \boldsymbol{x})^2(\overline{\boldsymbol{w}^\top \boldsymbol{y}})^2\right]^p \tag{20}$$

$$= \left(\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{l=1}^{d} \mathbb{E}[w_i w_j \overline{w_k w_l}] x_i x_j y_k y_l\right)^p \tag{21}$$

$\mathbb{E}_{ij\overline{kl}} := \mathbb{E}[w_i w_j \overline{w_k w_l}] \neq 0$, only if:

1. $i = j = k = l$: there are d terms $(\mathbb{E}_{ij\overline{kl}}) x_i x_j y_k y_l = \mathbb{E}[|w_i|^4] x_i^2 y_i^2$.
2. $i = k \neq j = l$: there are $d(d-1)$ terms $(\mathbb{E}_{ij\overline{kl}}) x_i x_j y_k y_l = \mathbb{E}[|w_i|^2] x_i y_i \mathbb{E}[|w_j|^2] x_j y_j = x_i y_i x_j y_j$.
3. $i = l \neq j = k$: there are $d(d-1)$ terms $(\mathbb{E}_{ij\overline{kl}}) x_i x_j y_k y_l = \mathbb{E}[|w_i|^2] x_i y_i \mathbb{E}[|w_j|^2] x_j y_j = x_i y_i x_j y_j$.

As for both the Gaussian and the Rademacher sketch, we have $\mathbb{E}[|w_i|^2] = 1$ for all $\{w_i\}_{i=1}^d$, we obtain:

$$\mathbb{E}\left[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})^2\right] = \left(\sum_{i=1}^{d} \mathbb{E}[|w_i|^4] x_i^2 y_i^2 + 2 \sum_{i=1}^{d} \sum_{j \neq i}^{d} x_i y_i x_j y_j\right)^p \tag{22}$$

We have $\mathbb{E}[|w_i|^4] = 2$ and $\mathbb{E}[|w_i|^4] = 1$ for the Gaussian and Rademacher case, respectively. So the pseudo-variances $\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] = \mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})^2] - \mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]^2$ are given by the following real-valued expressions:

$$\mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] = \frac{1}{D}\left(\left(2(\boldsymbol{x}^\top \boldsymbol{y})^2\right)^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p}\right) \qquad \text{(Gaussian)} \tag{23}$$

$$\mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] = \frac{1}{D}\left(\left(2(\boldsymbol{x}^\top \boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p}\right) \qquad \text{(Rademacher)} \tag{24}$$

where we added the $1/D$ scaling that we left out before. Note that $\mathbb{E}[|w_i|^4] \geq (\mathbb{E}[|w_i|^2])^2 = 1$ by Jensen's inequality, which is why the Rademacher sketch yields the lowest possible pseudo-variance for the estimator studied in Section 2.1.

Variance We work out $\mathbb{E}[|\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})|^2]$ to derive the variance $\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]$.

$$\begin{aligned}\mathbb{E}[|\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})|^2] &= \mathbb{E}\left[\left|\prod_{i=1}^{p} \boldsymbol{w}_i^\top \boldsymbol{x} \overline{\boldsymbol{w}_i^\top \boldsymbol{y}}\right|^2\right] = \mathbb{E}\left[\prod_{i=1}^{p} |\boldsymbol{w}_i^\top \boldsymbol{x}|^2 |\overline{\boldsymbol{w}_i^\top \boldsymbol{y}}|^2\right] \\ &= \mathbb{E}\left[(\sum_{i=1}^{d} w_i x_i)(\sum_{j=1}^{d} \overline{w_j} y_j)(\sum_{k=1}^{d} \overline{w_k} x_k)(\sum_{l=1}^{d} w_l y_l)\right]^p = \left(\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{l=1}^{d} \mathbb{E}[w_i \overline{w_j w_k} w_l] x_i y_j x_k y_l\right)^p.\end{aligned} \tag{25}$$

Now we check when $\mathbb{E}_{i\overline{jk}l} := \mathbb{E}\left[w_i \overline{w_j w_k} w_l\right] \neq 0$ holds. The analysis is the same as before with differently placed conjugates leading to different expressions.

1. $i = j = k = l$: there are d terms $(\mathbb{E}_{i\overline{jk}l}) x_i y_j x_k y_l = \mathbb{E}[|w_i|^4] x_i^2 y_i^2$.
2. $i = j \neq k = l$: there are $d(d-1)$ terms $(\mathbb{E}_{i\overline{jk}l}) x_i y_j x_k y_l = \mathbb{E}[|w_i|^2] \mathbb{E}[|w_k|^2] x_i x_k y_i y_k$.
3. $i = k \neq j = l$, there are $d(d-1)$ terms $(\mathbb{E}_{i\overline{jk}l}) x_i y_j x_k y_l = \mathbb{E}[|w_i|^2] \mathbb{E}[|w_j|^2] x_i^2 y_j^2$.
4. $i = l \neq j = k$, there are $d(d-1)$ terms $(\mathbb{E}_{i\overline{jk}l}) x_i y_j x_k y_l = \mathbb{E}[w_i^2] \mathbb{E}[\overline{w_j}^2] x_i x_j y_i y_j$.

Therefore,

$$\begin{aligned}\mathbb{E}[|\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})|^2]^{1/p} &= \sum_{i=1}^{d} \mathbb{E}[|w_i|^4] x_i^2 y_i^2 + \sum_{i=1}^{d} \sum_{\substack{j=1 \\ j \neq i}}^{d} x_i^2 y_j^2 + \sum_{i=1}^{d} \sum_{\substack{j=1 \\ j \neq i}}^{d} x_i x_j y_i y_j + \sum_{i=1}^{d} \sum_{\substack{j=1 \\ j \neq i}}^{d} \mathbb{E}[w_i^2] \mathbb{E}[\overline{w_j}^2] x_i x_j y_i y_j \\ &= \sum_{i=1}^{d} \mathbb{E}[|w_i|^4] x_i^2 y_i^2 + \left[\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right] + \left[(\boldsymbol{x}^\top \boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right] + \sum_{i=1}^{d} \sum_{\substack{j=1 \\ j \neq i}}^{d} \mathbb{E}[w_i^2] \mathbb{E}[\overline{w_j}^2] x_i x_j y_i y_j\end{aligned}$$

Once again, we have $\mathbb{E}[|w_i|^4] = 2$ and $\mathbb{E}[|w_i|^4] = 1$ for the Gaussian and Rademacher case, respectively. We further have $\mathbb{E}[w_i^2] = \mathbb{E}[\overline{w_i}^2] = \mathbb{E}[\mathrm{Re}\{w_i\}^2] - \mathbb{E}[\mathrm{Im}\{w_i\}^2]$ with $-1 \leq \mathbb{E}[w_i^2] \leq 1$ because $\mathbb{E}[|w_i|^2] = \mathbb{E}[\mathrm{Re}\{w_i\}^2] + \mathbb{E}[\mathrm{Im}\{w_i\}^2] = 1$.

Thus, $\mathbb{E}[w_i^2]\mathbb{E}[\overline{w_j}^2] = \mathbb{E}[w_i^2]^2 \in [0,1]$, where the extreme cases 0 and 1 are achieved by sampling w_i from $\{1, 1, \mathrm{i}, -\mathrm{i}\}$ (complex Rademacher) and $\{1, -1\}$ (real Rademacher), respectively. Therefore, we define the variable $q := (1 + \mathbb{E}[w_i^2]^2)$ that equals 1 for the complex case and 2 for the real one. We finally obtain the following variances $\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] = \mathbb{E}[|\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})|^2] - |\mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]|^2$:

$$\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] = \frac{1}{D}\left(\left(\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 + q(\boldsymbol{x}^\top \boldsymbol{y})^2\right)^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p}\right) \qquad \text{(Gaussian)} \tag{26}$$

$$\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] = \frac{1}{D}\left(\left(\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 + q\sum_{i=1}^{d}\sum_{j\neq i}^{d} x_i x_j y_i y_j\right)^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p}\right) \qquad \text{(Rademacher)} \tag{27}$$

where we added the $1/D$ scaling that we left out before. Note also that $\mathbb{E}[|w_i|^4] \geq (\mathbb{E}[|w_i|^2])^2 = 1$ by Jensen's inequality, which is why the (real/complex) Rademacher sketch yields the lowest possible variance for the estimator studied in Section 2.1.

Thus, when $\sum_{i=1}^{d}\sum_{j\neq i}^{d} x_i x_j y_i y_j \geq 0$, sampling w_i uniformly from $\{1, -1, \mathrm{i}, -\mathrm{i}\}$ yields the lowest possible CtR-variances as both the variance as well as the pseudo-variance lower bound are attained. In the opposite case, real Rademacher sketches (sampling w_i from $\{1, -1\}$) yield the lowest variances. This is because $\mathbb{E}[|\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})|^2] \geq 0$ is minimized in this case.

B.3 Gaussian and Rademacher CtR Variance Advantage over their Real-Valued Analogs

In the following, we compare Gaussian and Rademacher CtR-sketches against their real-valued analogs assuming that the corresponding feature maps have equal dimensions. Thus, we assign D random features to the real feature map $\Phi_{\mathrm{R}} : \mathbb{R}^d \to \mathbb{R}^D$ and only $D/2$ random feature samples to the CtR feature map $\Phi_{\mathrm{CtR}} : \mathbb{R}^d \to \mathbb{R}^D$ (Alg. 1) leading to the same output dimension D.

We call the corresponding kernel estimates $\hat{k}_{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y}) = \Phi_{\mathrm{R}}(\boldsymbol{x})^\top \Phi_{\mathrm{R}}(\boldsymbol{y})$ and $\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y}) = \Phi_{\mathrm{CtR}}(\boldsymbol{x})^\top \Phi_{\mathrm{CtR}}(\boldsymbol{y})$. $\mathbb{V}[\hat{k}_{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y})]$ is given in Eq. 26 and 27, where we set $q = 2$.

We further have $\mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})] = \frac{1}{2}(\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})] + \mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})])$ as shown in Section B.1. $\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]$ is given in Eq. 26 and 27, where we set $q = 1$. $\mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x}, \boldsymbol{y})]$ is given in Eq. 23 and 24, respectively.

We start with the simpler Gaussian case and study the Rademacher case after.

B.3.1 Gaussian Case: Proof of Theorem 3.5.

Proof Taking into account that $\boldsymbol{S}_{\mathrm{CtR}}$ has only $D/2$ rows for Φ_{R} and Φ_{CtR} to have equal dimensions D, the variance difference of their kernel estimates yields:

$$\begin{aligned}
&\mathbb{V}[\hat{k}_{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y})] - \mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})] \\
&= \frac{1}{D}\left(\left(\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 + 2(\boldsymbol{x}^\top \boldsymbol{y})^2\right)^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p}\right) - \frac{1}{D}\left(\left(\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 + (\boldsymbol{x}^\top \boldsymbol{y})^2\right)^p + \left(2(\boldsymbol{x}^\top \boldsymbol{y})^2\right)^p - 2(\boldsymbol{x}^\top \boldsymbol{y})^{2p}\right) \\
&= \frac{1}{D}\left(\left(\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 + 2(\boldsymbol{x}^\top \boldsymbol{y})^2\right)^p - \left(2(\boldsymbol{x}^\top \boldsymbol{y})^2\right)^p\right) - \frac{1}{D}\left(\left(\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 + (\boldsymbol{x}^\top \boldsymbol{y})^2\right)^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p}\right) \\
&= \frac{1}{D}\sum_{k=0}^{p-1}\binom{p}{k}\left(2(\boldsymbol{x}^\top \boldsymbol{y})^2\right)^k \left(\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2\right)^{p-k} - \frac{1}{D}\sum_{k=0}^{p-1}\binom{p}{k}(\boldsymbol{x}^\top \boldsymbol{y})^{2k}\left(\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2\right)^{p-k} \\
&= \frac{1}{D}\sum_{k=0}^{p-1}\binom{p}{k}(2^k - 1)(\boldsymbol{x}^\top \boldsymbol{y})^{2k}\left(\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2\right)^{p-k} \geq 0
\end{aligned}$$

■

Thus, the Gaussian CtR-estimator is always better regardless of the choice of $\boldsymbol{x}, \boldsymbol{y}$ and p and despite using only half the random feature samples. Note that the variance difference is zero if $p = 1$ and increases as p increases. Moreover, the difference is maximized for parallel $\boldsymbol{x}$ and $\boldsymbol{y}$. In this case, we have $(\boldsymbol{x}^\top \boldsymbol{y}) = \|\boldsymbol{x}\| \|\boldsymbol{y}\|$ and the difference becomes

$$\mathbb{V}[\hat{k}_{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y})] - \mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x}, \boldsymbol{y})] = \frac{1}{D}\sum_{k=0}^{p-1}\binom{p}{k}(2^k - 1)\left(\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2\right)^p = \frac{1}{D}\|\boldsymbol{x}\|^{2p} \|\boldsymbol{y}\|^{2p}\left(3^p - 2^{p+1} + 1\right)$$

We analyze the more difficult Rademacher case next.

B.3.2 Rademacher Case: Proof of Theorem 3.4.

Proof Taking into account that $\boldsymbol{S}_{\mathrm{CtR}}$ has only $D/2$ rows for Φ_{R} and Φ_{CtR} to have equal dimensions D, the variance difference of their kernel estimates yields:

$$\begin{aligned}
&\mathbb{V}[\hat{k}_{\mathrm{R}}(\boldsymbol{x},\boldsymbol{y})] - \mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x},\boldsymbol{y})] \\
&= \frac{1}{D}\left(\left(\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2 + 2\sum_{i=1}^{d}\sum_{j\neq i} x_i x_j y_i y_j\right)^p - (\boldsymbol{x}^\top\boldsymbol{y})^{2p}\right) \\
&\quad - \frac{1}{D}\left\{\left(\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2 + \sum_{i=1}^{d}\sum_{j\neq i} x_i x_j y_i y_j\right)^p - (\boldsymbol{x}^\top\boldsymbol{y})^{2p} + \left(2(\boldsymbol{x}^\top\boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^p - (\boldsymbol{x}^\top\boldsymbol{y})^{2p}\right\}
\end{aligned}$$

Next, we write $(\boldsymbol{x}^\top\boldsymbol{y})^{2p} = ((\boldsymbol{x}^\top\boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2 + \sum_{i=1}^{d} x_i^2 y_i^2)^p$. In this way, we can factor out the term $a := (\boldsymbol{x}^\top\boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2 = \sum_{i=1}^{d}\sum_{j\neq i} x_i x_j y_i y_j$ and apply the binomial theorem to *all* addends. This gives:

$$\begin{aligned}
\mathbb{V}[\hat{k}_{\mathrm{R}}(\boldsymbol{x},\boldsymbol{y})] - \mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x},\boldsymbol{y})] &= \frac{1}{2D}\sum_{k=0}^{p}\binom{p}{k} a^{p-k} \\
&\left(\left(\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2 + (\boldsymbol{x}^\top\boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^k - \left((\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2)^k + (\boldsymbol{x}^\top\boldsymbol{y})^{2k} - (\sum_{i=1}^{d} x_i^2 y_i^2)^k\right)\right)
\end{aligned}$$

We now show that the following term is always non-negative:

$$B := \left(\left(\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2 + (\boldsymbol{x}^\top\boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^k - \left((\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2)^k + (\boldsymbol{x}^\top\boldsymbol{y})^{2k} - (\sum_{i=1}^{d} x_i^2 y_i^2)^k\right)\right) \tag{28}$$

For $k = 0$ and $k = 1$, $B = 0$. For $k \geq 2$, we have:

$$\left(\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2 + (\boldsymbol{x}^\top\boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^k = \sum_{j=0}^{k}\binom{k}{j}\|\boldsymbol{x}\|^{2j}\|\boldsymbol{y}\|^{2j}\left((\boldsymbol{x}^\top\boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^{k-j}$$

Plugging this expression into B and cancelling out the addend for $j = k$ yields:

$$B = \sum_{j=0}^{k-1}\binom{k}{j}\|\boldsymbol{x}\|^{2j}\|\boldsymbol{y}\|^{2j}\left((\boldsymbol{x}^\top\boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2\right)^{k-j} - \left((\boldsymbol{x}^\top\boldsymbol{y})^{2k} - (\sum_{i=1}^{d} x_i^2 y_i^2)^k\right)$$

Next, we refactor $(\boldsymbol{x}^\top\boldsymbol{y})^{2k} - (\sum_{i=1}^{d} x_i^2 y_i^2)^k$:

$$\begin{aligned}
(\boldsymbol{x}^\top\boldsymbol{y})^{2k} - (\sum_{i=1}^{d} x_i^2 y_i^2)^k &= ((\boldsymbol{x}^\top\boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2 + \sum_{i=1}^{d} x_i^2 y_i^2)^k - (\sum_{i=1}^{d} x_i^2 y_i^2)^k \\
&= \sum_{j=0}^{k}\binom{k}{j}(\sum_{i=1}^{d} x_i^2 y_i^2)^j((\boldsymbol{x}^\top\boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2)^{k-j} - (\sum_{i=1}^{d} x_i^2 y_i^2)^k \\
&= \sum_{j=0}^{k-1}\binom{k}{j}(\sum_{i=1}^{d} x_i^2 y_i^2)^j((\boldsymbol{x}^\top\boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2)^{k-j}
\end{aligned}$$

Plugging this expression into B yields:

$$B = \sum_{j=0}^{k-1}\binom{k}{j}\left(\|\boldsymbol{x}\|^{2j}\|\boldsymbol{y}\|^{2j} - (\sum_{i=1}^{d} x_i^2 y_i^2)^j\right)((\boldsymbol{x}^\top\boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2)^{k-j}$$

Finally, we insert B back into the original variance difference $\mathbb{V}[\hat{k}_{\mathrm{R}}(\boldsymbol{x},\boldsymbol{y})] - \mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x},\boldsymbol{y})]$ (remember that $B = 0$ if $k < 2$):

$$\mathbb{V}[\hat{k}_{\mathrm{R}}(\boldsymbol{x},\boldsymbol{y})] - \mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x},\boldsymbol{y})] = \frac{1}{D}\sum_{k=2}^{p}\binom{p}{k}a^{p-k}B = \frac{1}{D}\sum_{k=2}^{p}\sum_{j=0}^{k-1}\binom{p}{k}\binom{k}{j}a^{p-j}\left(\|\boldsymbol{x}\|^{2j}\|\boldsymbol{y}\|^{2j} - (\sum_{i=1}^{d}x_i^2y_i^2)^j\right)$$

Finally, we note that $b_j := \|\boldsymbol{x}\|^{2j}\|\boldsymbol{y}\|^{2j} - (\sum_{i=1}^{d}x_i^2y_i^2)^j = (\sum_{i=1}^{d}\sum_{\ell=1}^{d}x_i^2y_\ell^2)^j - (\sum_{i=1}^{d}x_i^2y_i^2)^j \geq 0$ and $\mathbb{V}[\hat{k}_{\mathrm{R}}(\boldsymbol{x},\boldsymbol{y})] - \mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x},\boldsymbol{y})] \geq 0$ if $a = \sum_{i=1}^{d}\sum_{j'\neq i}^{d}x_ix_{j'}y_iy_{j'} \geq 0$. ■

C VARIANCE OF OUR PROPOSED (CtR-)ProductSRHT SKETCH

In Section 4, we proposed a novel (CtR-) ProductSRHT sketch that is a slightly modified version of the TensorSRHT sketch proposed by Ahle et al. (2020). Unlike previous work, we derive the variance of (CtR-)ProductSRHT and show its statistical advantage over unstructured sketches. The statistical advantage stems from the orthogonality of $\boldsymbol{H}$ as well as from the sampling matrices $\{\boldsymbol{P}_i\}_{i=1}^{p}$ that sample *without replacement*. The statistical advantage is lost when sampling *with replacement* as is done in Ahle et al. (2020). In this case, the variance falls back to the Rademacher variance.

C.1 Variances of ProductSRHT as well as CtR-ProductSRHT

As shown in Section B.1, the variance of the CtR sketches discussed in this work is of the form:

$$\mathbb{V}[\hat{k}_{\mathrm{CtR}}(\boldsymbol{x},\boldsymbol{y})] = \frac{1}{2}(\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})] + \mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})]),$$

where $\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})$ is the complex-valued kernel estimate of the polynomial kernel obtained through our sketch. In order to derive the variance of CtR-ProductSRHT, we need to derive the variance $\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})]$ and the pseudo-variance $\mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})]$. We will also derive the variance of real-valued ProductSRHT as a corollary of the variance of complex ProductSRHT.

C.1.1 Pseudo-variance

As before, we start with the pseudo-variance and derive the variance after. For the pseudo-variance $\mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})] = \mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})^2] - \mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})]^2$, we need to work out $\mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})^2]$:

$$\mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})^2] = \frac{1}{D^2}\sum_{\ell=1}^{D}\sum_{\ell'=1}^{D}\prod_{i=1}^{p}\mathbb{E}\left[(\boldsymbol{w}_{i,\ell}^\top\boldsymbol{x})(\overline{\boldsymbol{w}_{i,\ell}^\top\boldsymbol{y}})(\boldsymbol{w}_{i,\ell'}^\top\boldsymbol{x})(\overline{\boldsymbol{w}_{i,\ell'}^\top\boldsymbol{y}})\right] = \frac{1}{D^2}\sum_{\ell=1}^{D}\sum_{\ell'=1}^{D}\underbrace{\mathbb{E}\left[(\boldsymbol{w}_{\ell}^\top\boldsymbol{x})(\overline{\boldsymbol{w}_{\ell}^\top\boldsymbol{y}})(\boldsymbol{w}_{\ell'}^\top\boldsymbol{x})(\overline{\boldsymbol{w}_{\ell'}^\top\boldsymbol{y}})\right]^p}_{e(\ell,\ell')^p}$$

We dropped the index i in the last equality for ease of notation, as all $\{\boldsymbol{w}_{i,\ell}\}_{i=1}^{p}$ are i.i.d. samples and the expectation is thus the same for any i. To work out the expectation $e(\ell,\ell')$, we need to distinguish different cases for ℓ and ℓ'.

1. $\ell = \ell'$ (D terms): $e(\ell,\ell')^p = \mathbb{E}\left[(\boldsymbol{w}_\ell^\top\boldsymbol{x})^2(\overline{\boldsymbol{w}_\ell^\top\boldsymbol{y}})^2\right]^p = \left(2(\boldsymbol{x}^\top\boldsymbol{y}^2) - \sum_{i=1}^{d}x_i^2y_i^2\right)^p$
   (taken from Eq. 22 for the Rademacher case)

2. $\ell \neq \ell'$ ($D(D-1)$ terms):

$$\begin{aligned}e(\ell,\ell')^p &= \left(\sum_{q=1}^{d}\sum_{r=1}^{d}\sum_{s=1}^{d}\sum_{t=1}^{d}\mathbb{E}[w_{\ell,q}\overline{w_{\ell,r}}w_{\ell',s}\overline{w_{\ell',t}}]x_qy_rx_sy_t\right)^p \\ &= \left(\sum_{q=1}^{d}\sum_{r=1}^{d}\sum_{s=1}^{d}\sum_{t=1}^{d}\mathbb{E}[d_q\overline{d_r}d_s\overline{d_t}]\mathbb{E}[h_{p_\ell,q}h_{p_\ell,r}h_{p_{\ell'},s}h_{p_{\ell'},t}]x_qy_rx_sy_t\right)^p\end{aligned}$$

d_q, d_r, d_s, d_t are uniform samples from $\{1, -1, \mathrm{i}, -\mathrm{i}\}$, i.e., complex Rademacher samples, that are independent from the index samples $p_{\ell,q}, p_{\ell,r}, p_{\ell',s}, p_{\ell',t}$, which is why we can factor out the two expectations. We will simplify the above sum by studying when $\mathbb{E}[d_q\overline{d_r}d_s\overline{d_t}] \neq 0$.

We have to distinguish three non-zero cases for $\mathbb{E}[d_q\overline{d_r}d_s\overline{d_t}]$:

1. $q = r = s = t$ (d terms): $\mathbb{E}[d_q\overline{d_r}d_s\overline{d_t}] = \mathbb{E}[|d_q|^4] = 1$
2. $q = r \neq s = t$ ($d(d-1)$ terms): $\mathbb{E}[d_q\overline{d_r}d_s\overline{d_t}] = \mathbb{E}[|d_q|^2]\mathbb{E}[|d_s|^2] = 1$
3. $q = t \neq r = s$ ($d(d-1)$ terms): $\mathbb{E}[d_q\overline{d_r}d_s\overline{d_t}] = \mathbb{E}[|d_q|^2]\mathbb{E}[|d_r|^2] = 1$

because $\mathbb{E}[|d_q|^4] = \mathbb{E}[|d_q|^2] = 1$.

In Section C.1.3, we show that for $\ell \neq \ell'$ and $q \neq r$, $\mathbb{E}[h_{p_\ell,q}h_{p_\ell,r}h_{p_{\ell'},r}h_{p_{\ell'},q}] = -\frac{1}{\lceil D/d \rceil d - 1}$ holds. Therefore, $e(\ell,\ell')^p$ for $\ell \neq \ell'$ yields:

$$
\begin{aligned}
e(\ell,\ell')^p &= \left(\sum_{i=1}^{d} x_i^2 y_i^2 + \sum_{i=1}^{d}\sum_{j\neq i}^{d} x_i y_i x_j y_j - \frac{1}{\lceil D/d \rceil d - 1} \sum_{i=1}^{d}\sum_{j\neq i}^{d} x_i y_i x_j y_j \right)^p \\
&= \left((\boldsymbol{x}^\top \boldsymbol{y})^2 - \frac{1}{\lceil D/d \rceil d - 1} \left[(\boldsymbol{x}^\top \boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2 \right] \right)^p
\end{aligned}
$$

In fact, $e(\ell,\ell')^p$ does not depend on ℓ and ℓ' anymore after working out the expectations involved. Plugging $e(\ell,\ell')^p$ back into $\mathbb{E}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})^2]$ yields the following pseudo-variance for ProductSRHT:

$$
\begin{aligned}
\mathbb{PV}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})^2] &= \frac{1}{D}\left[\left(2(\boldsymbol{x}^\top \boldsymbol{y}^2) - \sum_{i=1}^{d} x_i^2 y_i^2 \right)^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p} \right] \\
&\quad + \left(1 - \frac{1}{D}\right)\left[\left((\boldsymbol{x}^\top \boldsymbol{y})^2 - \frac{1}{\lceil D/d \rceil d - 1} \left[(\boldsymbol{x}^\top \boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2 \right] \right)^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p} \right] \\
&= \frac{1}{D}\mathbb{PV}_{\mathrm{Rad.}}^{(p)} - \left(1 - \frac{1}{D}\right)\left[(\boldsymbol{x}^\top \boldsymbol{y})^{2p} - \left((\boldsymbol{x}^\top \boldsymbol{y})^2 - \frac{\mathbb{PV}_{\mathrm{Rad.}}^{(1)}}{\lceil D/d \rceil d - 1} \right)^p \right]
\end{aligned} \tag{29}
$$

$\mathbb{PV}_{\mathrm{Rad.}}^{(p)}$ and $\mathbb{PV}_{\mathrm{Rad.}}^{(1)}$ are the Rademacher pseudo-variance (24) for a given degree p and $p = 1$, respectively.

C.1.2 Variance

Next we work out the variance $\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})]$:

$$
\mathbb{V}\left[\frac{1}{D} \sum_{\ell=1}^{D} \prod_{i=1}^{p} (\boldsymbol{w}_{i,\ell}^\top \boldsymbol{x})(\overline{\boldsymbol{w}_{i,\ell}^\top \boldsymbol{y}}) \right] = \frac{1}{D^2} \sum_{\ell=1}^{D}\sum_{\ell'=1}^{D} \mathrm{Cov}\left(\prod_{i=1}^{p} (\boldsymbol{w}_{i,\ell}^\top \boldsymbol{x})(\overline{\boldsymbol{w}_{i,\ell}^\top \boldsymbol{y}}), \prod_{i=1}^{p} (\boldsymbol{w}_{i,\ell'}^\top \boldsymbol{x})(\overline{\boldsymbol{w}_{i,\ell'}^\top \boldsymbol{y}}) \right)
$$

Again, we distinguish the cases $\ell = \ell'$ and $\ell \neq \ell'$:

1. $\ell = \ell'$ (D terms):

$$
\begin{aligned}
&\mathrm{Cov}\left(\prod_{i=1}^{p} (\boldsymbol{w}_{i,\ell}^\top \boldsymbol{x})(\overline{\boldsymbol{w}_{i,\ell}^\top \boldsymbol{y}}), \prod_{i=1}^{p} (\boldsymbol{w}_{i,\ell}^\top \boldsymbol{x})(\overline{\boldsymbol{w}_{i,\ell}^\top \boldsymbol{y}}) \right) = \mathbb{V}\left[\prod_{i=1}^{p} (\boldsymbol{w}_{i,\ell}^\top \boldsymbol{x})(\overline{\boldsymbol{w}_{i,\ell}^\top \boldsymbol{y}}) \right] \\
&= \left(\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 + (\boldsymbol{x}^\top \boldsymbol{y})^2 - \sum_{i=1}^{d} x_i^2 y_i^2 \right)^p - (\boldsymbol{x}^\top \boldsymbol{y})^{2p} \quad \text{(Using the complex Rademacher variance (27))}
\end{aligned}
$$

2. $\ell \neq \ell'$ ($D(D-1)$ terms). We discuss this case in detail below.

$$\operatorname{Cov}\left(\prod_{i=1}^{p}(\boldsymbol{w}_{i,\ell}^\top\boldsymbol{x})(\overline{\boldsymbol{w}_{i,\ell}^\top\boldsymbol{y}}),\prod_{i=1}^{p}(\boldsymbol{w}_{i,\ell'}^\top\boldsymbol{x})(\overline{\boldsymbol{w}_{i,\ell'}^\top\boldsymbol{y}})\right)=\mathbb{E}\left[\prod_{i=1}^{p}(\boldsymbol{w}_{i,\ell}^\top\boldsymbol{x})(\overline{\boldsymbol{w}_{i,\ell}^\top\boldsymbol{y}})(\overline{\boldsymbol{w}_{i,\ell'}^\top\boldsymbol{x}})(\overline{\boldsymbol{w}_{i,\ell'}^\top\boldsymbol{y}})\right]-(\boldsymbol{x}^\top\boldsymbol{y})^{2p} \tag{30}$$
$$=\mathbb{E}\left[(\boldsymbol{w}_\ell^\top\boldsymbol{x})(\overline{\boldsymbol{w}_\ell^\top\boldsymbol{y}})(\overline{\boldsymbol{w}_{\ell'}^\top\boldsymbol{x}})(\overline{\boldsymbol{w}_{\ell'}^\top\boldsymbol{y}})\right]^p-(\boldsymbol{x}^\top\boldsymbol{y})^{2p}=\underbrace{\mathbb{E}\left[(\boldsymbol{w}_\ell^\top\boldsymbol{x})(\overline{\boldsymbol{w}_\ell^\top\boldsymbol{y}})(\overline{\boldsymbol{w}_{\ell'}^\top\boldsymbol{x}})(\boldsymbol{w}_{\ell'}^\top\boldsymbol{y})\right]^p}_{e_2(\ell,\ell')^p}-(\boldsymbol{x}^\top\boldsymbol{y})^{2p}$$

Next, we turn to the expression $e_2(\ell,\ell')^p$ that is almost the same as $e(\ell,\ell')^p$ for the pseudo-variance, the only difference being the complex conjugates that are placed differently:

$$\begin{aligned}e_2(\ell,\ell')^p&=\left(\sum_{q=1}^{d}\sum_{r=1}^{d}\sum_{s=1}^{d}\sum_{t=1}^{d}\mathbb{E}[w_{\ell,q}\overline{w_{\ell,r}w_{\ell',s}}w_{\ell',t}]x_qy_rx_sy_t\right)^p\\&=\left(\sum_{q=1}^{d}\sum_{r=1}^{d}\sum_{s=1}^{d}\sum_{t=1}^{d}\mathbb{E}[d_q\overline{d_rd_s}d_t]\mathbb{E}[h_{p_\ell,q}h_{p_\ell,r}h_{p_{\ell'},s}h_{p_{\ell'},t}]x_qy_rx_sy_t\right)^p\end{aligned}$$

We distinguish 4 cases for $\mathbb{E}[d_q\overline{d_rd_s}d_t]$:

1. $q=r=s=t$ (d terms): $\mathbb{E}[d_q\overline{d_rd_s}d_t]=\mathbb{E}[|d_q|^4]=1$
2. $q=r\neq s=t$ ($d(d-1)$ terms): $\mathbb{E}[d_q\overline{d_rd_s}d_t]=\mathbb{E}[|d_q|^2]\mathbb{E}[|d_s|^2]=\mathbb{E}[|d_q|^2]^2=1$
3. $q=s\neq r=t$ ($d(d-1)$ terms): $\mathbb{E}[d_q\overline{d_rd_s}d_t]=\mathbb{E}[|d_q|^2]\mathbb{E}[|d_r|^2]=\mathbb{E}[|d_q|^2]^2=1$
4. $q=t\neq r=s$ ($d(d-1)$ terms): $\mathbb{E}[d_q\overline{d_rd_s}d_t]=\mathbb{E}[d_q^2]\mathbb{E}[\overline{d_r}^2]=0$

We showed case (4) on purpose although it is zero for complex Rademacher samples $d_q,d_r\in\mathbb{C}$. For real Rademacher samples, we have $\mathbb{E}[d_q^2]=\mathbb{E}[\overline{d_r}^2]=1$ instead. This observation will allow us to work out the variance of complex and real ProductSRHT at the same time. Furthermore, we have $\mathbb{E}[h_{p_\ell,q}h_{p_\ell,r}h_{p_{\ell'},r}h_{p_{\ell'},q}]=-\frac{1}{\lceil D/d\rceil d}$ for any $q\neq r$ and $\ell\neq\ell'$ as already noted for the pseudo-variance. The derivation of this quantity is shown in Section C.1.3.

So $e_2(\ell,\ell')$ reduces to:

$$\begin{aligned}e_2(\ell,\ell')&=\underbrace{\sum_{i=1}^{d}x_i^2y_i^2}_{\text{Case (1)}}+\underbrace{\sum_{i=1}^{d}\sum_{j\neq i}^{d}x_ix_jy_iy_j}_{\text{Case (2)}}-\underbrace{\frac{1}{\lceil D/d\rceil d}\sum_{i=1}^{d}\sum_{j\neq i}^{d}x_i^2y_j^2}_{\text{Case (3)}}-\underbrace{\frac{1}{\lceil D/d\rceil d}\sum_{i=1}^{d}\sum_{j\neq i}^{d}\mathbb{E}[d_i^2]\mathbb{E}[\overline{d_j}^2]x_ix_jy_iy_j}_{\text{Case (4)}}\\&=(\boldsymbol{x}^\top\boldsymbol{y})^2-\frac{1}{\lceil D/d\rceil d}\sum_{i=1}^{d}\sum_{j\neq i}^{d}x_i^2y_j^2+\mathbb{E}[d_i^2]x_ix_jy_iy_j,\end{aligned}$$

where $\mathbb{E}[d_i^2]=0$ for the complex case and $\mathbb{E}[d_i^2]=1$ for the real case. Plugging back $e_2(\ell,\ell')$ for the case $\ell\neq\ell'$ back into Eq. 30 and solving for $\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})]$ yields:

$$\mathbb{V}[\hat{k}_{\mathrm{C}}(\boldsymbol{x},\boldsymbol{y})]=\mathbb{V}_{\mathrm{Rad.}}^{(p)}-\left(1-\frac{1}{D}\right)\left[(\boldsymbol{x}^\top\boldsymbol{y})^{2p}-\left((\boldsymbol{x}^\top\boldsymbol{y})^2-\frac{\mathbb{V}_{\mathrm{Rad.}}^{(1)}}{\lceil D/d\rceil d-1}\right)^p\right] \tag{31}$$

with $\mathbb{V}_{\mathrm{Rad.}}^{(p)}$ and $\mathbb{V}_{\mathrm{Rad.}}^{(1)}$ being the Rademacher variance (27) for a given degree p and $p=1$, respectively. We set $q=2$ for the real case and $q=1$ for the complex case inside Eq. 27.

Inserting the expressions for the variance (31) and pseudo-variance (29) into Eq. 7, gives the variance of CtR-ProductSRHT.

C.1.3 Shuffling the Rows of Stacked Hadamard Matrices

In this section, we prove an important equality that was used in the derivation of the variance formulas of ProductSRHT in the previous sections. It can be seen as the key lemma that leads to a reduced variance compared to Rademacher sketches.

It shows the statistics of randomly sampled rows (without replacement) inside stacked orthogonal Hadamard matrices that give close-to-orthogonal as opposed to i.i.d. samples in our proposed ProductSRHT sketch. We prove the equality

$$\mathbb{E}[h_{p_\ell,q}h_{p_\ell,r}h_{p_{\ell'},r}h_{p_{\ell'},q}] = -\frac{1}{\lceil D/d\rceil d - 1}$$

for $\ell \neq \ell'$ and $q \neq r$ being fixed indices. $\boldsymbol{h}_{p_\ell}^\top$ and $\boldsymbol{h}_{p_{\ell'}}^\top$ are the p_ℓ-th and $p_{\ell'}$-th row of the Hadamard matrix $\boldsymbol{H}$, respectively (see Section 4). The indices q and r refer to elements inside these row vectors. p_ℓ and $p_{\ell'}$ are themselves the $\pi(\ell)$-th and $\pi(\ell')$-th entries of the vector $\boldsymbol{p}_i \in \mathbb{R}^{\lceil D/d\rceil d}$ for a given $i \in \{1,\dots,p\}$. Here, we look at a given index i and drop the index for ease of presentation. We do the same for the permutation function $\pi(\cdot)$. Recall that $\{\boldsymbol{p}_i\}_{i=1}^p$ is used to construct the sampling matrices $\{\boldsymbol{P}_i\}_{i=1}^p$ in Alg. 2.

The following proof is closely related to Choromanski et al. (2017, Proof of Proposition 8.2) and Wacker et al. (2022, Lemma B.1). The difference here is that we consider the sampling of rows (without replacement) inside *stacked* Hadamard matrices as we will see next, whereas the other works only consider the sampling of rows inside a *single* Hadamard matrix.

Proof

The sampling procedure for the rows $\boldsymbol{h}_{p_\ell}^\top$ and $\boldsymbol{h}_{p_{\ell'}}^\top$ can be described as follows. We stack the Hadamard matrix $\boldsymbol{H} \in \mathbb{R}^{d\times d}$ $\lceil D/d\rceil$ times on top of itself to yield a new matrix $\boldsymbol{H}^{\lceil D/d\rceil} \in \mathbb{R}^{\lceil D/d\rceil d\times d}$. We then shuffle its rows randomly to yield the shuffled matrix $\boldsymbol{H}_{\boldsymbol{p}}^{\lceil D/d\rceil\times d}$. $\boldsymbol{h}_{p_\ell}^\top$ and $\boldsymbol{h}_{p_{\ell'}}^\top$ are then the ℓ-th and ℓ'-th row of $\boldsymbol{H}_{\boldsymbol{p}}^{\lceil D/d\rceil}$. In fact, the shuffled matrix $\boldsymbol{H}_{\boldsymbol{p}}^{\lceil D/d\rceil}$ can be constructed from the index vector $\boldsymbol{p}$ that contains the order of the rows of $\boldsymbol{H}$ to be used.

Since the columns of $\boldsymbol{H}$ are orthogonal, the same is true for $\boldsymbol{H}^{\lceil D/d\rceil}$ and $\boldsymbol{H}_{\boldsymbol{p}}^{\lceil D/d\rceil}$. So the inner product of two distinct columns q and r of $\boldsymbol{H}_{\boldsymbol{p}}^{\lceil D/d\rceil}$ yields $\sum_{\ell=1}^{\lceil D/d\rceil d} h_{p_\ell,q}h_{p_\ell,r} = 0$. As $h_{p_\ell,q}, h_{p_\ell,r} \in \{1,-1\}$, half of $\{h_{p_\ell,q}h_{p_\ell,r}\}_{\ell=1}^{\lceil D/d\rceil d}$ must be equal to 1 and -1, respectively. From this we get the marginal probabilities

$$\Pr(h_{p_\ell,q}h_{p_\ell,r} = 1) = \Pr(h_{p_\ell,q}h_{p_\ell,r} = -1) = 0.5$$

for any $q \neq r$ being fixed, where the probabilities are taken over the indices p_ℓ and p_ℓ, i.e., the shuffling operation. Next, we obtain the following conditional probabilities using the same logic as before:

$$\Pr(h_{\ell',q}h_{\ell',r} = 1 | h_{\ell,q}h_{\ell,r} = 1) = \Pr(h_{\ell',q}h_{\ell',r} = -1 | h_{\ell,q}h_{\ell,r} = -1) = \frac{(\lceil D/d\rceil d)/2 - 1}{\lceil D/d\rceil d - 1}$$
$$\Pr(h_{\ell',q}h_{\ell',r} = 1 | h_{\ell,q}h_{\ell,r} = -1) = \Pr(h_{\ell',q}h_{\ell',r} = -1 | h_{\ell,q}h_{\ell,r} = 1) = \frac{(\lceil D/d\rceil d)/2}{\lceil D/d\rceil d - 1}$$

Using these conditional probabilities along with the marginal probabilities $\Pr(h_{p_\ell,q}h_{p_\ell,r})$ allows us to solve $\mathbb{E}[h_{p_\ell,q}h_{p_\ell,r}h_{p_{\ell'},r}h_{p_{\ell'},q}]$ via the law of total expectation:

$$\begin{aligned}
\mathbb{E}[h_{p_\ell,q}h_{p_\ell,r}h_{p_{\ell'},r}h_{p_{\ell'},q}] &= \mathbb{E}_{p_\ell}[\mathbb{E}_{p_{\ell'}}[h_{p_\ell,q}h_{p_\ell,r}h_{p_{\ell'},r}h_{p_{\ell'},q}|h_{p_\ell,r}h_{p_\ell,q}]] \\
&= \frac{1}{2}\left(\mathbb{E}_{p_{\ell'}}[h_{p_{\ell'},r}h_{p_{\ell'},q}|h_{p_\ell,r}h_{p_\ell,q} = 1] - \mathbb{E}_{p_{\ell'}}[h_{p_{\ell'},r}h_{p_{\ell'},q}|h_{p_\ell,r}h_{p_\ell,q} = -1]\right) \\
&= \frac{1}{2}\left(\left(\frac{(\lceil D/d\rceil d)/2 - 1}{\lceil D/d\rceil d - 1} - \frac{(\lceil D/d\rceil d)/2}{\lceil D/d\rceil d - 1}\right) - \left(\frac{(\lceil D/d\rceil d)/2}{\lceil D/d\rceil d - 1} - \frac{(\lceil D/d\rceil d)/2 - 1}{\lceil D/d\rceil d - 1}\right)\right) \\
&= -\frac{1}{\lceil D/d\rceil d - 1}
\end{aligned}$$

■

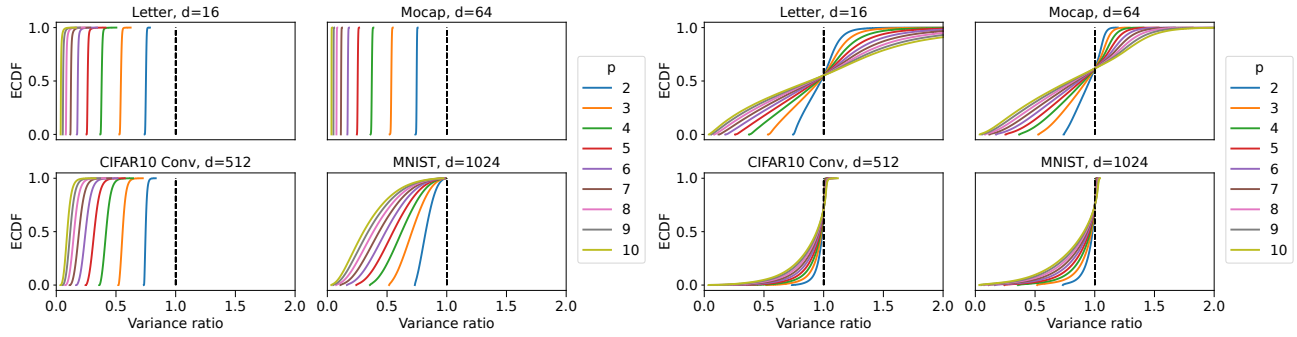
D FURTHER EXPERIMENTS

In this section, we provide further experiments complementing our evaluation in Section 6 of the main paper.

D.1 Empirical Variance Comparison of (CtR-) Rademacher Sketches

We first study the practical effect of the non-negativity condition $a = \sum_{i=1}^{d} \sum_{j' \neq i}^{d} x_i x_{j'} y_i y_{j'} \geq 0$ in Thm. 3.4. Fig. 6 shows the results of an empirical variance comparison of CtR-Rademacher sketches against their real analogs.

Fig. 6a shows the case, where the condition $a \geq 0$ always holds (non-negative data) and Fig. 6b the case, where $a \geq 0$ does not always hold (zero-centered data). While the CtR sketch offers lower variance ratios for CIFAR-10 and MNIST in most cases even if $a \geq 0$ does not always hold, we see that $a \geq 0$ is needed to *guarantee* an advantage of the CtR sketch. For Letter and Mocap with zero-centered data (Fig. 6b), around half the variances ratios are less than one and half are more than one, suggesting that real Rademacher sketches perform similarly to CtR-Rademacher sketches in this case. For non-negative data (Fig. 6a), the relative gains of CtR-sketches improve drastically. That is, all variance ratios are less than one, with an increasing gain for larger p.



(a) Non-negative data.

(b) Zero-centered data.

Figure 6: ECDF of Var(CtR-Rademacher) / Var(Rademacher) for pairwise evaluations of the variance ratio evaluated on a subset of each dataset.

D.2 Closed-Form GP Classification

We carry out a set of additional GP classification experiments to complement Section 6.2. The experiments are the same as in Section 6.2, but compare a larger range of values for D and two additional data sets: Letter and Mocap (Dua & Graff, 2017). Moreover, we add experiments for zero-centered data. The following is a brief summary of the plots:

- Fig. 8 shows MNIST/CIFAR-10 experiments for $p = 3, 7$ using unit-normalized non-negative data (same as Fig. 2 for a larger range of D).
- Fig. 9 shows Letter/Mocap experiments for $p = 3, 7$ using unit-normalized non-negative data.
- Fig. 10 shows MNIST/CIFAR-10 experiments for $p = 3, 7$ using unit-normalized zero-centered data.
- Fig. 11 shows Letter/Mocap experiments for $p = 3, 7$ using unit-normalized zero-centered data.

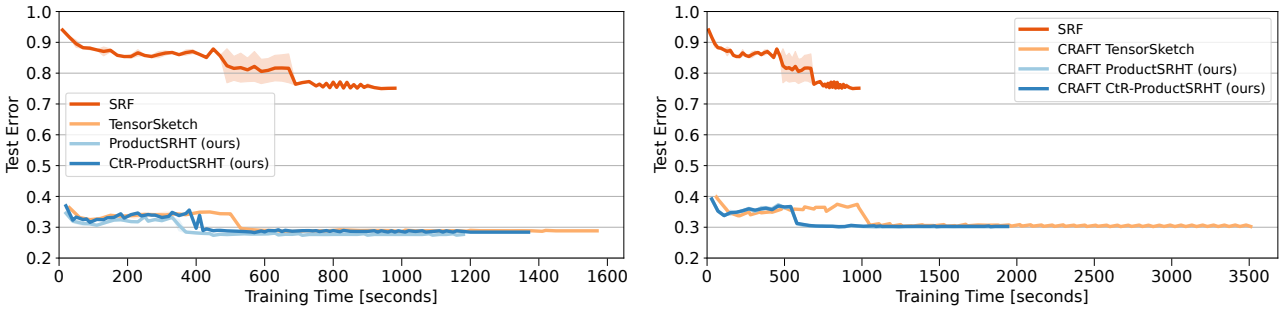
In general, we find that relative performance gains of CtR-sketches over their real analogs are larger for non-negative than for zero-centered data. This makes sense because of the condition of Thm. 3.4. However, they still lead to some improvements even for zero-centered data. Gains over SRF on the other hand increase for zero-centered data, in particular regarding kernel approximation errors.

D.3 Online Learning for Fine-Grained Visual Recognition

Fig. 7 shows an online learning experiment on the CUB-200 (Welinder et al., 2010) data set. We follow the experimental setup in Gao et al. (2016), but only train the classification layer of the VGG-M (Chatfield et al., 2014) convolutional neural network. This option is referred to as *no fine-tuning* in the original paper.

We use an Adam optimizer with decaying learning rate starting from 10^{-3}, where the learning rate is divided by 10, when the validation loss stagnates. The mini-batch size is 32 and we train over 50 epochs. The sketch dimension is $D = 2^{13}$ and $p = 3$, $a = 2$ in our experiments (see Section 6.1).

Our final test errors are lower than 36.42% and 31.53% for Rademacher and TensorSketch, respectively, given in Gao et al. (2016, Table 4). An exception is SRF that requires unit-normalized features and hence loses important information, leading to around 75% test error. Since the polynomial degree $p = 3$ is small, CtR-ProductSRHT does not achieve an advantage over ProductSRHT in terms of test errors. ProductSRHT is also slightly faster. When using CRAFT maps on the other hand, both CtR-ProductSRHT and ProductSRHT perform similarly well, and are significantly faster than TensorSketch.



(a) (CtR-) ProductSRHT vs. TensorSketch/SRF.

(b) Same as (a) with CRAFT maps.

Figure 7: Stochastic optimization following Gao et al. (2016) for the CUB-200 data set *without* fine-tuning of the VGG-M convolutional layers.

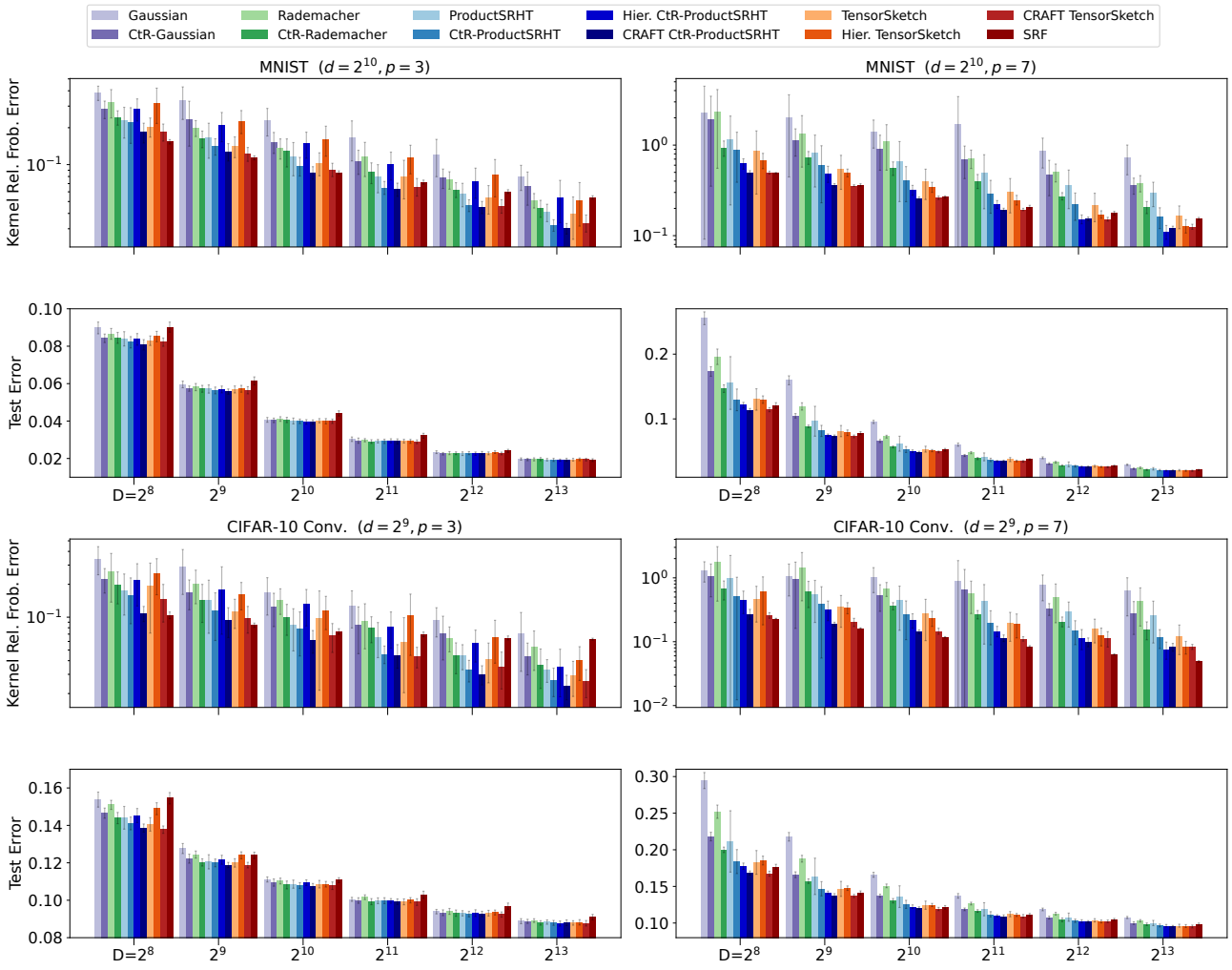


Figure 8: MNIST and CIFAR-10 comparison for $p = 3$ and $p = 7$ with unit-normalized data averaged over 20 seeds.

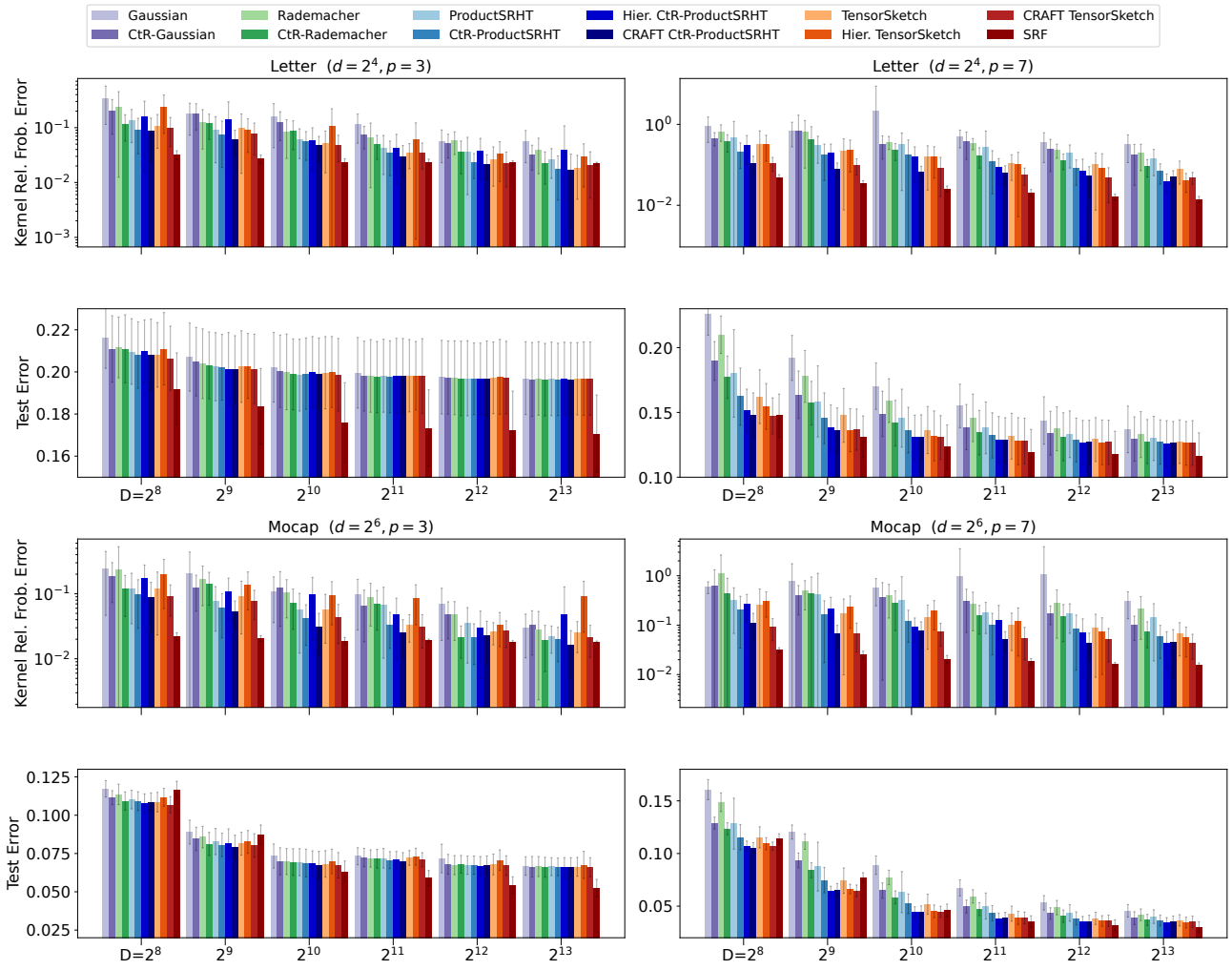


Figure 9: Letter and Mocap comparison for $p = 3$ and $p = 7$ with unit-normalized data averaged over 20 seeds.

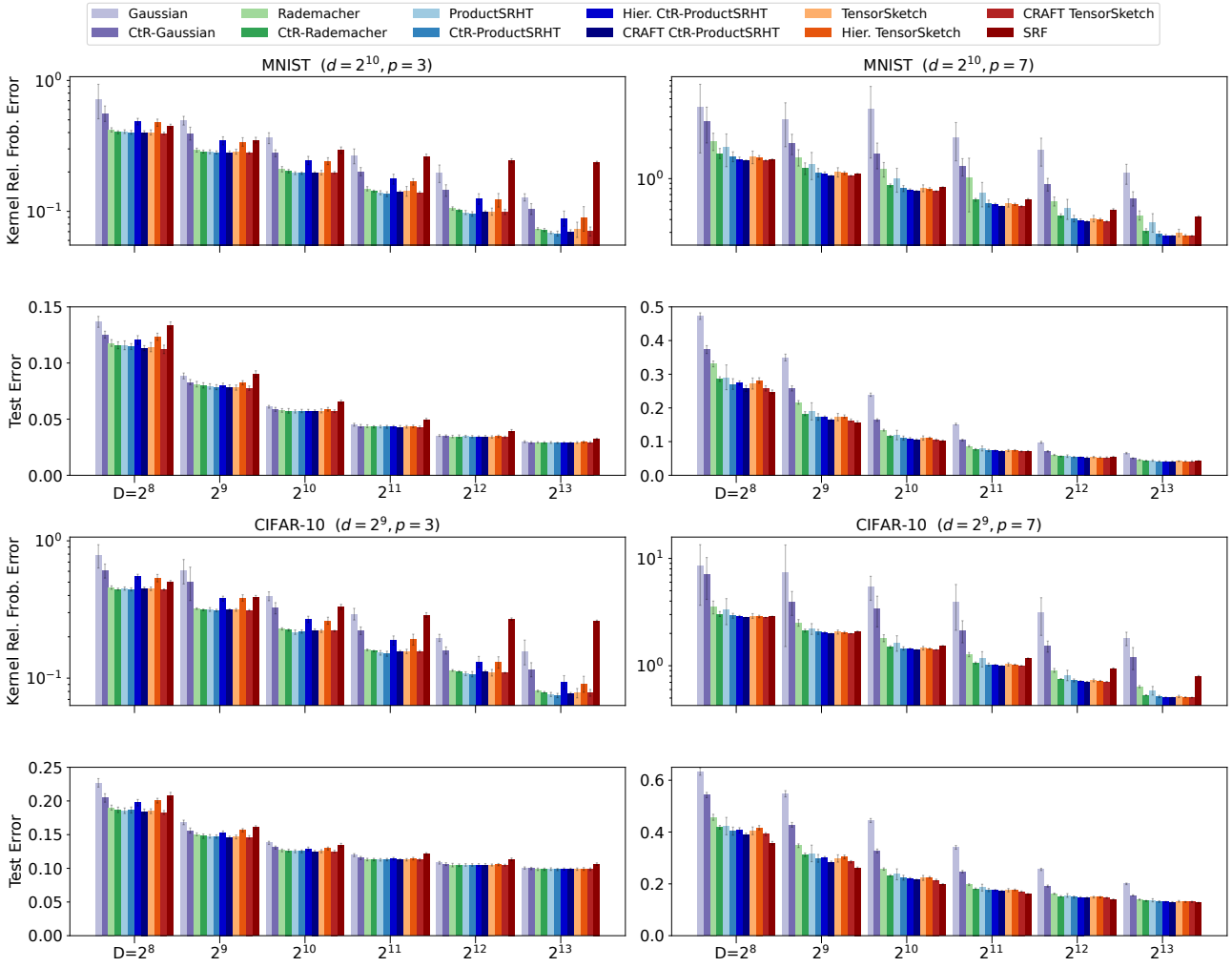


Figure 10: MNIST and CIFAR-10 comparison for $p = 3$ and $p = 7$ averaged over 20 seeds. The data is centered through a subtraction of the training mean and unit-normalized afterwards.

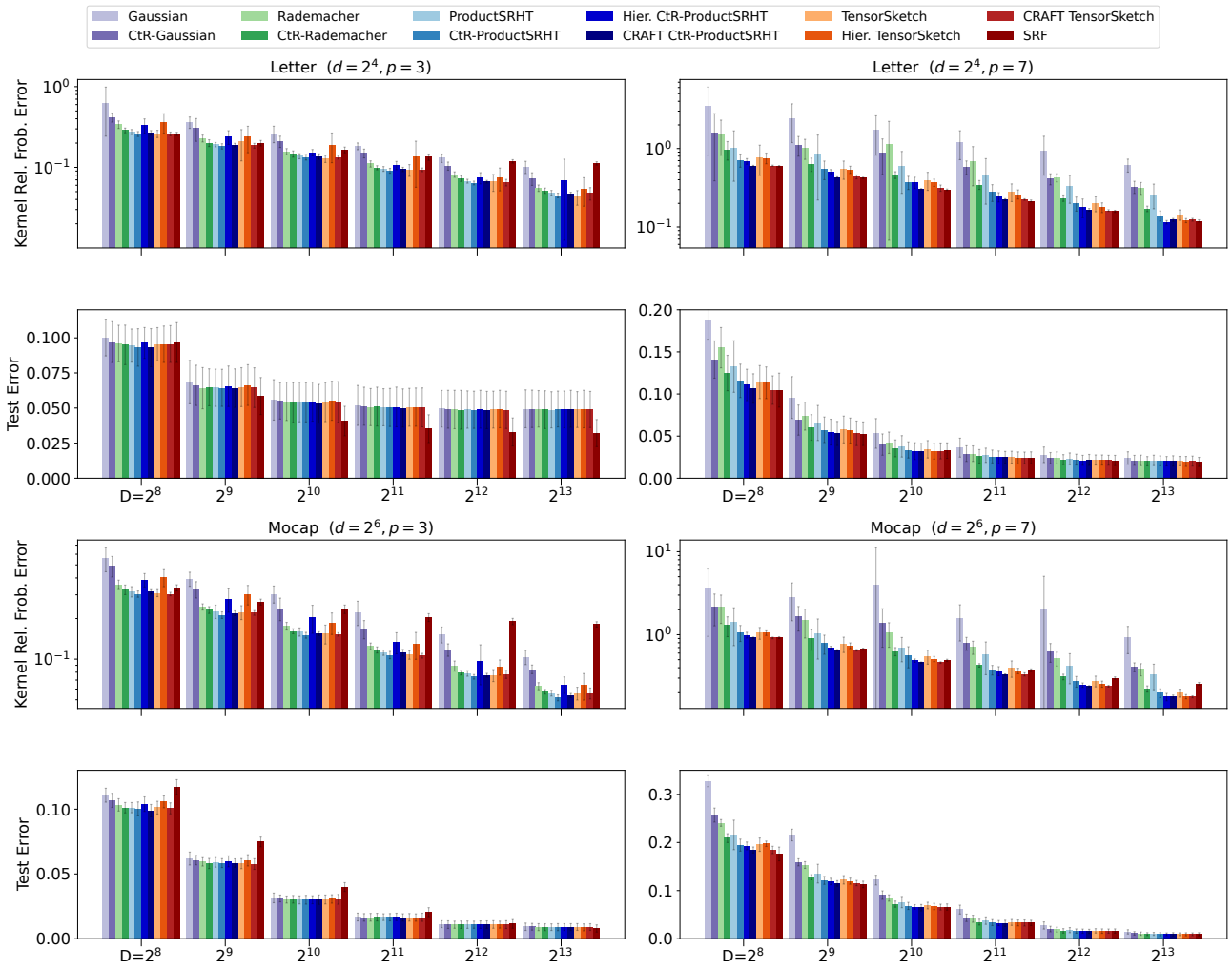


Figure 11: Letter and Mocap comparison for $p = 3$ and $p = 7$ averaged over 20 seeds. The data is centered through a subtraction of the training mean and unit-normalized afterwards.