RANDOM FOURIER FEATURES IN NEAR-LINEAR TIME

Anonymous authors
Paper under double-blind review

ABSTRACT

Random Fourier features approximate a shift-invariant kernel in $\mathbb{R}^d$, such as the Gaussian kernel $\exp(-\frac{1}{2}\|x-y\|_2^2)$, using independent samples from its spectral measure to construct an unbiased estimator for the kernel value with subgaussian concentration. However, generating m such features requires $O(md)$ time. A long line of work on structured random features has accelerated this computation for the Gaussian kernel and other radial kernels, albeit with weaker concentration guarantees. We close this gap by showing a random feature map that can be computed in near-linear time $O((d+m)\log d)$ and yields near-subgaussian concentration. We obtain it through a new and tighter analysis of the Fastfood feature map of Le, Sarlós and Smola (ICML 2013), using Brascamp-Lieb inequalities and the Schoenberg representation of radial kernels.

1 INTRODUCTION

A positive definite kernel $\mathrm{k}(x,y)$, such as the Gaussian kernel $e^{-\|x-y\|_2^2/2}$, can be viewed as an inner product $\mathrm{k}(x,y) = \phi(x)^T\phi(y)$ where $x \mapsto \phi(x)$ is a feature map into a space that may be infinite-dimensional. Kernel methods access this feature space implicitly through evaluations of k, without constructing the vectors $\phi(x)$ explicitly, an approach known as the "kernel trick". While widely useful, this flexibility can nevertheless be computationally costly: training over a dataset $\{x_i\}_{i=1}^n$ often requires storing or manipulating the $n \times n$ Gram matrix $(k(x_i,x_j))_{i,j}$, and evaluating a learned predictor on a test point y may require computing $\mathrm{k}(y,x_i)$ on many of the training points.

Random Fourier features (RFF) (Rahimi & Recht, 2007) address this problem for shift-invariant kernels by sampling m independent frequencies from the kernel's so called spectral measure (see Section 2) and replacing the kernel by an inner product in a $O(m)$-dimensional randomized feature space. That is, we approximate $\mathrm{k}(x,y)$ as the inner product $F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y)$ where F_{k} is called the (randomized) feature map and maps to $O(m)$ dimensions. For every fixed pair x, y, this estimator is unbiased and admits strong (subgaussian) concentration as m grows. RFF has become a cornerstone method in machine learning and continues to be widely used in modern AI systems and architectures (e.g., Peng et al. (2021); Saito et al. (2025); Nguyen et al. (2026); Zhang & Zhou (2026)). However, its drawback is that computing $F_{\mathrm{k}}(x)$ generally takes $O(md)$ time, which is a substantial computational bottleneck when the dimension d is large.

As a result, a long line of work has been dedicated to faster variants of RFF. Most work has focused on the Gaussian and other radial kernels through the use of structured transforms, mainly randomized Hadamard transforms (RHTs). An RHT is a linear map $x \mapsto HDx$, where H is the Hadamard matrix and D is a random diagonal matrix. An RHT can be computed in time $O(d\log d)$, while at the same time it mimics certain properties of a truly random rotation, which would have taken time $O(d^2)$ to compute. This makes it useful in efficient methods for rotation-invariant kernels (i.e., radial). Many fast RFF variants have been designed by simulating pseudo-random Fourier frequencies via repeated RHTs (Le et al., 2013; Yu et al., 2016; Choromanski & Sindhwani, 2016; Choromanski et al., 2017; Cherapanamjeri & Nelson, 2022; Backurs et al., 2024; Wagner, 2026).

This line of work has achieved a sequence of either provably or heuristically faster variants of RFF. However, it has not achieved an RFF variant with nearly linear time $\widetilde{O}(d+m)$ and nearly subgaussian concentration, leaving open the question whether such random features are possible.

Our results. We answer this question affirmatively by proving the existence of an unbiased random feature map with time $O((d+m)\log d)$ and strong concentration, for the Gaussian kernel and other kernels. Our random feature map is the simplified form of the Fastfood map due to Le et al. (2013),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Random feature maps $F : \mathbb{R}^d \to \mathbb{R}^m$ for the Gaussian kernel. The table lists the output dimension $m = m(\varepsilon, \delta)$ needed for the tail bound $\Pr[|\exp(-\frac{1}{2}\|x-y\|_2^2) - F(x)^T F(y)| \geq \varepsilon] \leq \delta$ for every fixed pair $x, y \in \mathbb{R}^d$, as well as the time it takes to compute $x \mapsto F(x)$. The dependence on δ is omitted since it is $\log^{O(1)}(1/\delta)$ in all the bounds. Δ denotes an upper bound on any $\|x-y\|_2$.

REFERENCE	METHOD	OUT DIM.	TIME FOR $x \mapsto F(x)$
Rahimi & Recht (2007)	Random Fourier features	$O(1/\varepsilon^2)$	$O(d/\varepsilon^2)$
Le et al. (2013)	Fastfood	$\widetilde{O}(\Delta^2/\varepsilon^2)$	$\widetilde{O}(d + \Delta^2/\varepsilon^2)$
Backurs et al. (2024)	Fast dim. reduction + RFF	$\widetilde{O}(1/\varepsilon^2)$	$\widetilde{O}(d + 1/\varepsilon^4)$
Wagner (2026)	Fast spherical emb. + Fastfood	$\widetilde{O}(1/\varepsilon^3)$	$\widetilde{O}(d + \varepsilon\Delta^2 + 1/\varepsilon^3)$
This work (Thm 1.1)	Fastfood (new analysis)	$\widetilde{O}(1/\varepsilon^2)$	$\widetilde{O}(d + 1/\varepsilon^2)$

and our contribution is a new and tighter analysis of it. Indeed, it was noted in Le et al. (2013) that the authors believed their analysis of Fastfood was pessimistic, and that a better tail bound could be provable through a different approach. Our work shows that that postulation was correct. The comparison between their bounds and ours, as well as other bounds, is shown in Table 1.

Kernel density estimation (KDE). KDE is a canonical use case for random features. The problem is to preprocess a finite set $X \subset \mathbb{R}^d$ such that given any fixed query $y \in \mathbb{R}^d$, we could report an estimate of $\frac{1}{|X|}\sum_{x \in X} \mathrm{k}(x, y)$ as quickly as possible. We consider *additive* error up to $\pm\varepsilon$, with high probability. There is a very simple reduction from KDE to random features. Given an unbiased and strongly concentrated feature map $F : \mathbb{R}^d \to \mathbb{R}^m$, we compute $F(X) = \frac{1}{|X|}\sum_{x \in X} F(x)$ during preprocessing, and then report $F(X)^T F(y)$ given any y. The query time is $O(m)$ plus the time to compute $y \mapsto F(y)$ (i.e., the sum of the two rightmost columns in Table 1). Our results yield the first KDE solution with query time $\widetilde{O}(d + 1/\varepsilon^2)$ for this additive approximation guarantee, answering affirmatively the main open question posed in Wagner (2026). See Section A.7.

Bandwidth. Kernels usually have a scaling parameter σ called bandwidth (e.g., the Gaussian kernel is $\exp(-\|x-y\|_2^2/(2\sigma^2))$. As a convention, we set $\sigma = 1$ throughout. This does not limit generality, since all the kernels we discuss are functions of the Euclidean distance between the input pair of points. Therefore, our results extend to any bandwidth by appropriately scaling the input points.

Limitations. As our contribution is a new and tighter analysis of the classical Fastfood method, our work does not introduce new algorithms. We do not include an experimental evaluation since the empirical performance of Fastfood is already comprehensively studied (Le et al., 2013; Yu et al., 2016; Feng et al., 2015; Choromanski & Sindhwani, 2016; Liu et al., 2021).

1.1 FORMAL STATEMENT OF RESULTS

Our main theorem is for the Gaussian kernel.

Theorem 1.1. *Let $d, m \geq 1$ be integers. There is a randomized map $F : \mathbb{R}^d \to \mathbb{R}^{2m}$, samplable in time $O(d+m)$ and computable in time $O((d+m)\log d)$, such that for every $x, y \in \mathbb{R}^d$:*

$$\mathbb{E}[F(x)^T F(y)] = \exp(-\|x-y\|_2^2/2), \tag{1}$$

and for all $t > 0$ and $0 < \eta < 1$:

$$\Pr\left[|F(x)^T F(y) - \exp(-\|x-y\|_2^2/2)| \geq t\right] \leq \eta + 2\exp\left(-\frac{mt^2}{4\log(2d/\eta)}\right). \tag{2}$$

Our next theorem extends Theorem 1.1 beyond the Gaussian kernel to other radial kernels under mild assumptions. A kernel $\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called *radial* if it depends only on the Euclidean distance between its inputs, that is, if there exists a map $\kappa : \mathbb{R} \to \mathbb{R}$ such that $\mathrm{k}(x-y) = \kappa(\|x-y\|_2)$ for every $x, y \in \mathbb{R}^d$. The kernel is called *normalized* if $\kappa(0) = 1$. We also need the kernel to be positive definite, which is defined formally in Section 2.

Theorem 1.2. *Let $d, m \geq 1$ be integers where d is a power of 2. Let $\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^d \to [-1, 1]$ be a continuous positive definite normalized radial kernel. There is a randomized map $F_{\mathrm{k}} : \mathbb{R}^d \to \mathbb{R}^{2m}$, computable in time $O((d+m)\log d)$, such that for every $x, y \in \mathbb{R}^d$:*

$$\mathbb{E}[F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y)] = \mathrm{k}(x, y), \tag{3}$$

and for all $t > 0$ and $0 < \eta < 1$:

$$\Pr\left[|F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y) - \mathrm{k}(x, y)| \geq t\right] \leq \eta + 2\exp\left(-\frac{mt^2}{32\log(2d/\eta)}\right) + 2\exp\left(-\frac{mt^2}{8}\right). \tag{4}$$

The time to sample F is $O(d + m(1 + T_{\mathrm{k}}^{\mathrm{rad}})$, where $T_{\mathrm{k}}^{\mathrm{rad}}$ is the time to draw a sample from the radial measure of k.

The radial measure of the kernel is the radial part of its spectral measure. We review the spectral measure in Section 2 and define the radial measure formally in Section 4. We show in Section A.6 that for many standard kernels, one can sample from the radial measure in expected time $O(1)$.

Lastly, we show that the running time in Theorem 1.1 is nearly optimal.

Theorem 1.3. *Let $d \geq 1$, $0 < \varepsilon \leq 1/8$, and $0 < \delta \leq \varepsilon^4$. Suppose a randomized map $F : \mathbb{R}^d \to \mathbb{R}^m$ satisfies, for every fixed x, y,*

$$\Pr\left[\left|\langle F(x), F(y)\rangle - \exp(-\|x - y\|_2^2/2)\right| > \varepsilon\right] \leq \delta. \tag{5}$$

Then, evaluating F must take time $\Omega(d + \frac{\log(1/\delta))}{\varepsilon^2 \log(1/\varepsilon)})$.

If we set $t = \varepsilon$, $\eta = \frac{1}{2}\delta$, $m = O(\frac{\log(1/\delta))\log(d/\delta)}{\varepsilon^2})$ in Theorem 1.1, Equation (2) becomes Equation (5). Thus, Theorem 1.3 shows that the computation time of F in Theorem 1.1 is optimal up to log factors.

1.2 Techniques

We present a new analysis for the Fastfood transform (Le et al., 2013). Fastfood is a sequence of two RHTs. The inner RHT acts as flattening on the input vector (i.e., spreading its ℓ_2-mass approximately equally across its coordinates), as per a classical result of Ailon & Chazelle (2009) (see Lemma 3.1). The challenge in analyzing the outer RHT is that, unlike original RFF, its output coordinates are not independent, thus requiring tail bound techniques for dependent random variables. Previous methods used concentration theorems for Lipschitz functions (Le et al., 2013) or Wiener chaoses (Wagner, 2026) over the Gaussian diagonal entries of the outer RHT. Unfortunately, these techniques sustain a term that grows with the squared norm of the input vector, leading to the excess dependence on the diameter Δ as seen in Table 1.

The main new step in our proof of Theorem 1.1 is to control the outer RHT with the classic Brascamp-Lieb inequality (Brascamp & Lieb, 1976; Lieb, 1990). Roughly, it states that if our random variables arise as the projections of a Gaussian vector on m "sufficiently well-spread" unit directions $\{w_i\}_{i=1}^m$, then although they are dependent, they concentrate like $c^* m$ truly independent random variables. $c^* \in (0, 1)$ is a parameter that measures how well-spread the directions are (namely, $c^* \sum_i w_i w_i^T$ must be sub-isotropic; see Equation (10)). In our analysis, the output features are cast as the projections of the Gaussian diagonal of the outer RHT onto scaled Hadamard directions, and the flattening property of the inner RHT ensures that c^* is sufficiently large, i.e., the directions are well-spread.

Theorem 1.2 extends this approach beyond the Gaussian kernel by using Schoenberg's representation (Schoenberg, 1938a;b) of radial kernels as mixtures of spherical shell kernels (see Equation (19)). Then the above reasoning is applied to each shell kernel with a spherical variant of Brascamp-Lieb, where the projected vector is uniform over the unit sphere. The lower bound in Theorem 1.3 follows from standard arguments (e.g., Chen & Phillips (2017)) as a consequence of Alon's rank bound for perturbed identity matrices (Alon, 2009) by designing an input dataset where the pairwise kernel values mimick a gram matrix of nearly orthogonal vectors.

1.3 Related work

Following Rahimi & Recht (2007), the methods summarized in Table 1 have been proposed as faster variants of random Fourier features. The Fastfood method, key to our work, was introduced in Le

et al. (2013) and analyzed through tail bounds for Lipschitz functions of Gaussians. Backurs et al. (2024) proved new guarantees for the fast dimension reduction theorem of Ailon & Chazelle (2009) and used it as a pre-processing step for RFF. Wagner (2026) proved a fast version of a spherical embedding theorem due to Bartal et al. (2011) and used it as pre-processing for Fastfood. The bounds arising from these methods are incomparable and depend on the parameter regime. All are based on RHTs, popularized in the algorithmic literature by Sarlos (2006); Ailon & Chazelle (2009).

The above methods yield a "for-each" guarantee that holds for each fixed pair with high probability. Fast RFF with a stronger "for-all" guarantee, which holds simultaneously for all pairs in a bounded domain, was studied in Cherapanamjeri & Nelson (2022). In a different vein, Yu et al. (2016) proposed Structured Orthogonal Random Features (SORF), a fast variant of RFF that is also based on RHTs. Although highly effective empirically, it does not establish a kernel estimation tail bound comparable to the methods in Table 1.

In addition to RFF, other approximate kernel feature methods include Quasi Monte-Carlo (Avron et al., 2016; Huang et al., 2024), quadrature (Bach, 2017; Munkhoeva et al., 2018), Nyström (Rudi et al., 2015; Gittens & Mahoney, 2016), and hashing-based methods (Andoni & Indyk, 2009; Charikar & Siminelakis, 2017; Siminelakis et al., 2019; Backurs et al., 2019).

2 Random Fourier features: technical background

We review the basics of RFF due to Rahimi & Recht (2007). Let $\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a map that we refer to as a kernel over $\mathbb{R}^d$. The kernel is *shift-invariant* if $\mathrm{k}(x, y)$ depends only on $x - y$ for every $x, y \in \mathbb{R}^d$, that is, if $\mathrm{k}(x, y) = \mathrm{k}(x - y)$ with a slight abuse of notation. The kernel is *positive definite* if $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \mathrm{k}(x_i, x_j) \geq 0$ for every $n \in \mathbb{N}$, $c_1, \ldots, c_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathbb{R}^d$.

By the classical Bochner theorem, a continuous shift-invariant kernel satisfying $\mathrm{k}(0) = 1$ is positive definite if and only if it is the Fourier transform of a probability measure μ_{k} over $\mathbb{R}^d$:

$$\mathrm{k}(x - y) = \mathbb{E}_{\omega \sim \mu_{\mathrm{k}}}[e^{\mathbf{i} \cdot \omega^T (x-y)}], \tag{6}$$

where $\mathbf{i}$ denotes the imaginary unit. μ_{k} is unique and called the *spectral measure* of k. For example, for the Gaussian kernel $\mathrm{k}(x, y) = \exp(-\|x - y\|_2^2/2)$, the spectral measure is the standard multivariate Gaussian distribution, $N(0, I_d)$.

The goal of random features is to find a map $F_{\mathrm{k}} : \mathbb{R}^d \to \mathbb{R}^m$ such that $F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y)$ is a good approximation for $\mathrm{k}(x, y)$ for each pair $x, y \in \mathbb{R}^d$. Rahimi & Recht (2007) proposed to obtain this from Equation (6) via Monte-Carlo sampling. If one draws iid samples $\omega_1, \ldots, \omega_m$ from the spectral measure μ_{k}, and defines $F_{\mathrm{k}} : \mathbb{R}^d \to \mathbb{R}^{2m}$ by

$$F_{\mathrm{k}}(x) = \frac{1}{\sqrt{m}} \oplus_{i=1}^{m} \left(\sin(\omega_i^T x),\ \cos(\omega_i^T x) \right), \tag{7}$$

then from Equation (6) one can show that

$$\forall\ x, y \in \mathbb{R}^d \quad \mathrm{k}(x, y) = \mathbb{E}_{\{\omega_i\}}[F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y)]. \tag{8}$$

Furthermore, since the coordinates of F are iid and bounded in $[-1, 1]$, the standard Chernoff-Hoeffding tail bound yields strong concentration about the mean:

$$\forall\ x, y \in \mathbb{R}^d,\ \forall\ t > 0, \quad \Pr[\left| F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y) - \mathrm{k}(x, y) \right| \geq t] \leq 2 \exp(-mt^2/2). \tag{9}$$

Thus, once computed, random features can be used to speed kernel computations in a variety of downstream tasks. Unfortunately, computing $F_{\mathrm{k}}(x)$ takes time $O(md)$, which could on its own be computationally heavy, particularly where the dimension d is large. The goal of the present paper, and of the long line of work it belongs to, is to achieve properties similar to Equations (8) and (9) for natural kernels with a faster computation time.

Remark 2.1. *Another possible computational bottleneck is to sample F_{k} to begin with (before applying it to any point). This takes time $O(m(1 + T_{\mathrm{k}}^{\mathrm{spec}}))$, where $T_{\mathrm{k}}^{\mathrm{spec}}$ is the time to draw a single sample from μ_{k}. For many natural kernels, including the Gaussian kernel, $T_{\mathrm{k}}^{\mathrm{spec}} = \Theta(d)$.*

Remark 2.2. *Another qualitatively similar way to define RFF is to also draw uniform iid $\beta_1, \ldots, \beta_m \sim [0, 2\pi]$ and define $\check{F}_{\mathrm{k}}(x) = \sqrt{2m^{-1}} \oplus_{i=1}^{m} \cos(\omega_i^T x + \beta_i)$. Both forms of RFF appear widely in the literature. We focus on Equation (7) for simplicity, since our analysis comes out cleaner without the additional source of randomness $\{\beta_i\}$.*

3 FAST RANDOM FEATURES FOR THE GAUSSIAN KERNEL (THEOREM 1.1)

We assume that d is a power of 2 w.l.o.g. by padding with 0s. We also focus on the case $m \leq d$ for clarity since that is the challenging regime; the extension to $m > d$ then follows as detailed in Section A.4. Our map is the simplified form of the Fastfood map from Le et al. (2013). We denote:

$$V := \sqrt{d} HGHB,$$

where $H \in \mathbb{R}^{d\times d}$ is the normalized Hadamard matrix (its entries are $\pm 1/\sqrt{d}$), $B \in \mathbb{R}^{d\times d}$ is a diagonal matrix with iid random signs on the diagonal, and $G \in \mathbb{R}^{d\times d}$ is a random diagonal matrix with iid $N(0,1)$s on the diagonal. The map $F : \mathbb{R}^d \to \mathbb{R}^{2m}$ is defined as

$$F(x) = \frac{1}{\sqrt{m}} \oplus_{i=1}^{m} \left(\sin\left((Vx)_i\right), \cos\left((Vx)_i\right)\right).$$

3.1 PRELIMINARIES

The next lemmas are either standard or shown in prior work. We include details in the appendix.

Lemma 3.1 (Fast JL flattening (Ailon & Chazelle, 2009)). *For every $z \in \mathbb{R}^d$ and $\eta > 0$,*

$$\Pr_B\left[\|HBz\|_\infty \leq \|z\|_2 \sqrt{\frac{2\log(2d/\eta)}{d}}\right] > 1 - \eta.$$

Lemma 3.2 (see Section A.3). *For every $z \in \mathbb{R}^d$ and every supported B, $\|HBz\|_2 = \|z\|_2$.*

Lemma 3.3 (Le et al. (2013); see Section A.3). *For every $z \in \mathbb{R}^d$, conditioned on B, each entry of Vz is marginally distributed like $N(0, \|z\|_2^2)$.*

Lemma 3.4 (standard fact). $\mathbb{E}_{X\sim N(0,\nu^2)}[\cos(X)] = e^{-\nu^2/2}$.

Lemma 3.5 (Hoeffding's lemma). *Let X be a real-valued random variable such that $X \in [a, b]$ w.p. 1. Then for all $\lambda > 0$, $\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq e^{\lambda^2(b-a)^2/8}$.*

Lemma 3.6 (rank-1 Gaussian Brascamp-Lieb; see Section A.1). *Let $w_1, \ldots, w_k \in \mathbb{R}^d$ be unit vectors, let $c_1, \ldots, c_k > 0$, and suppose*

$$\sum_{i=1}^{k} c_i w_i w_i^T \preceq I_d. \tag{10}$$

Then, for all non-negative measurable $f_i : \mathbb{R} \to [0, \infty)$,

$$\mathbb{E}_{g\sim N(0,I_d)}\left[\prod_{i=1}^{k} f_i(w_i^T g)^{c_i}\right] \leq \prod_{i=1}^{k} \left(\mathbb{E}[f_i(N(0,1))]\right)^{c_i}. \tag{11}$$

3.2 PROOF OF THEOREM 1.1

Fix a pair $x, y \in \mathbb{R}^d$ and let $z := x - y$. We start by proving Equation (1). The following lemma follows from standard trigonometric identities, see Section A.3.

Lemma 3.7. $F(x)^T F(y) = \frac{1}{m}\sum_{i=1}^{m} \cos((Vz)_i)$.

By Lemma 3.3 $(Vz)_i \sim N(0, \|z\|_2^2)$ for every i, thus Equation (1) follows from Lemma 3.4.

We proceed to prove Equation (2). We first outline the argument and the use of Brascamp-Lieb (Lemma 3.6). Suppose $X_1, \ldots, X_m$ are zero-centered random variables and $X = \frac{1}{m}\sum_i X_i$. Our goal is to prove a tail bound of the form $\Pr[X \geq t] \leq f(m, t)$. The standard Chernoff-Hoeffding proof blueprint goes as follows: for every $s > 0$, we have $\Pr[X \geq t] = \Pr\left[e^{sX} \geq e^{st}\right] \leq e^{-st}\mathbb{E}\left[e^{sX}\right]$ by Markov's inequality. Thus the tail bound reduces to bounding the moment generating function (MGF) $\mathbb{E}[e^{sX}]$. In the simplest case, if $\{X_i\}$ are independent, we have

$$e^{-st}\mathbb{E}\left[e^{\frac{s}{m}\sum_i X_i}\right] = e^{-st}\mathbb{E}\left[\prod_i e^{\frac{s}{m}X_i}\right] = e^{-st}\prod_i \mathbb{E}\left[e^{\frac{s}{m}X_i}\right]. \tag{12}$$

Then, a tail bound can be obtained by bounding the MGFs of each individual X_i (often through Hoeffding's lemma) and optimizing for the choice of s that produces the tightest bound.

In our case, $\{X_i\}$ would be associated with the output features of the structured random feature map F in Theorem 1.1. Unfortunately, they are not independent, since every output feature depends on all the random diagonal entries of B and G. Fortunately, the Brascamp-Lieb inequality Lemma 3.6 provides an alternative way to bound the MGF, which can replace the rightmost inequality in Equation (12). It requires that the X_is are determined by the projection of a single standard Gaussian vector g onto directions $w_1, \ldots, w_m$, and that we can assign them weights $c_1, \ldots, c_m$ to render them sub-isotropic, meaning to satisfy Equation (10). Intuitively, this condition ensures that no direction is overrepresented, and $\{c_i\}$ quantify how overrepresented any direction in $\mathbb{R}^d$ could be by the collection $\{w_i\}$. With a common weight $c_i = c^*$, Brascamp-Lieb allows us to bound the MGF of the m terms as if they were approximately c^*m independent random variables.

This matches our setting because conditioned on B, every output feature of F is determined by $w_i^T g$ where g is the diagonal vector of G and the w_is are fixed directions. Namely, for a fixed $u \in \mathbb{R}^d$, the outer Hadamard stage, HG, satisfies $HGu = H\mathrm{diag}(u)g$. Since the rows of H are orthonormal, the sub-isotropy condition boils down to bounding the magnitude of entries in u. By Fast JL flattening (Lemma 3.1), the inner Hadamard stage, HB, ensures that $u = HBz$ is an orthogonal image of z that with high probability has entry magnitude nearly as small as possible. Taken together, the two Hadamard stages in F allow us to utilize Brascamp-Lieb instead of the fully independent factorization in the rightmost equality of Equation (12).

We now turn to the formal proof. We may assume that $z \neq 0$, since if $z = 0$ then $F(x)^T F(y) = 1$ by Lemma 3.7, rendering Equation (2) trivial. We let

$$X_i = \cos((Vz)_i) - \exp(-\|z\|_2^2/2),$$

and observe that by Lemma 3.7, $F(x)^T F(y) - \exp(-\|z\|_2^2/2) = \frac{1}{m}\sum_{i=1}^m X_i$. Conditioned on B, as per the foregoing discussion, for every $s > 0$ we have by Markov's inequality,

$$\Pr\left[\frac{1}{m}\sum_i X_i \geq t \mid B\right] = \Pr\left[e^{\frac{s}{m}\sum_i X_i} \geq e^{st} \mid B\right] \leq e^{-st}\mathbb{E}\left[\prod_i e^{\frac{s}{m}X_i} \mid B\right],$$

which, by writing out $\{X_i\}$, becomes

$$\begin{aligned}&\Pr\left[F(x)^T F(y) - \exp(-\|z\|_2^2/2) \geq t \mid B\right] \\ &\leq \exp(-st)\cdot \mathbb{E}\left[\prod_{i=1}^m \exp\left(\frac{s}{m}\left(\cos((Vz)_i) - \exp(-\|z\|_2^2/2)\right)\right) \mid B\right]. \end{aligned} \tag{13}$$

We now use Brascamp-Lieb (Lemma 3.6) to bound the right-hand side. Recall that $u = HBz$. Let $g = (g_1, \ldots, g_d) \sim N(0, I_d)$ be the diagonal entries of G. Observe that the ith entry of Vz can be written as

$$(Vz)_i = e_i^T Vz = e_i^T(\sqrt{d}HGu) = e_i^T(\sqrt{d}H\mathrm{diag}(g)u) = e_i^T(\sqrt{d}H\mathrm{diag}(u)g).$$

Thus, if we let

$$w_i = \frac{\sqrt{d}}{\|z\|_2}\mathrm{diag}(u)H^T e_i,$$

then

$$(Vz)_i = \|z\|_2 \cdot w_i^T g. \tag{14}$$

First, we verify that w_i has unit length. Since $\sqrt{d}H^T e_i$ is a vector with all entries in $\{\pm 1\}$,

$$\|w_i\|_2 = \frac{1}{\|z\|_2}\|\mathrm{diag}(u)\cdot\sqrt{d}H^T e_i\|_2 = \frac{\|u\|_2}{\|z\|_2} = 1,$$

having used Lemma 3.2. Next, we choose (c_i) to satisfy the sub-isotropy condition Equation (10). We have,

$$\sum_{i=1}^m w_i w_i^T \leq \sum_{i=1}^d w_i w_i^T = \frac{d}{\|z\|_2^2}\sum_{i=1}^d \mathrm{diag}(u)H^T e_i e_i^T H\mathrm{diag}(u)$$

$$= \frac{d}{\|z\|_2^2}\mathrm{diag}(u)H^T\left(\sum_{i=1}^{d} e_i e_i^T\right)H\mathrm{diag}(u) = \frac{d}{\|z\|_2^2}\cdot\mathrm{diag}(u^2) \preceq \frac{d\|u\|_\infty^2}{\|z\|_2^2}\cdot I_d.$$

Therefore, letting each c_i be equal to

$$c^* = \frac{\|z\|_2^2}{d\|u\|_\infty^2}, \tag{15}$$

we have Equation (10). Finally, we define

$$f_i(\alpha) = \exp\left(\frac{s}{c^*\cdot m}\left(\cos(\|z\|_2\cdot\alpha) - \exp\left(-\|z\|_2^2/2\right)\right)\right).$$

Then by Equation (14),

$$f_i(w_i^T g)^{c^*} = \exp\left(\frac{s}{m}\left(\cos((Vz)_i) - \exp\left(-\|z\|_2^2/2\right)\right)\right).$$

This completes the setup for Brascamp-Lieb. Lemma 3.6 yields

$$\mathbb{E}\left[\prod_{i=1}^{m}\exp\left(\frac{s}{m}\left(\cos((Vz)_i) - \exp(-\|z\|_2^2/2)\right)\right) \mid B\right] \le \prod_{i=1}^{m}\left(\mathbb{E}[f_i(N(0,1))]\right)^{c^*}. \tag{16}$$

To control each individual $\mathbb{E}[f_i(N(0,1))]$, let Z be the random variable $Z = \cos(\|z\|_2\cdot N(0,1))$. It is clearly bounded in $[-1,1]$ w.p. 1, and by Lemma 3.4, $\mathbb{E}[Z] = \exp\left(-\|z\|_2^2/2\right)$. Therefore, by Hoeffding's lemma (Lemma 3.5),

$$\begin{aligned}\mathbb{E}[f_i(N(0,1))] &= \mathbb{E}\left[\exp\left(\frac{s}{c^*\cdot m}\left(\cos(\|z\|_2\cdot N(0,1)) - \exp\left(-\|z\|_2^2/2\right)\right)\right)\right]\\ &= \mathbb{E}\left[\exp\left(\frac{s}{c^*\cdot m}\left(Z - \mathbb{E}[Z]\right)\right)\right] \le \exp\left(\frac{s^2}{2(c^*)^2m^2}\right).\end{aligned}$$

Plugging this in Equation (16) yields

$$\mathbb{E}\left[\prod_{i=1}^{m}\exp\left(\frac{s}{m}\left(\cos((Vz)_i) - \exp(-\|z\|_2^2/2)\right)\right) \mid B\right] \le \exp\left(\frac{s^2}{2c^*m}\right). \tag{17}$$

Plugging into Equation (13) and choosing $s = tc^*m$, we get

$$\Pr\left[F(x)^T F(y) - \exp(-\|z\|_2^2/2) \ge t \mid B\right] \le \exp\left(-t^2c^*m/2\right).$$

The proof of the lower tail is similar, thus

$$\Pr\left[|F(x)^T F(y) - \exp(-\|z\|_2^2/2)| \ge t \mid B\right] \le 2\exp\left(-t^2c^*m/2\right). \tag{18}$$

Finally, we bound c^* by un-conditioning on B. By the Fast JL flattening lemma (Lemma 3.1), we have $\|u\|_\infty^2 \le \|z\|_2^2\cdot\frac{2\log(2d/\eta)}{d}$ w.p. $1-\eta$ over the random choice of B. Thus, by Equation (15), $c^* \ge 1/(2\log(2d/\eta))$. Plugging this in Equation (18) with a union bound over the failure of the flattening event (losing an additive η to the overall failure probability on the right-hand side), we get

$$\Pr\left[|F(x)^T F(y) - \exp(-\|z\|_2^2/2)| \ge t\right] \le \eta + 2\exp\left(-t^2m/(4\log(2d/\eta))\right),$$

which is Equation (2). This concludes the proof of Theorem 1.1.

4 Fast random features for other radial kernels (Theorem 1.2)

We first define the radial measure of a kernel.

Definition 4.1 (radial measure of a kernel). *Let* $\mathrm{k}:\mathbb{R}^d\times\mathbb{R}^d\to[-1,1]$ *be a shift-invariant positive definite kernel, and let* μ_k *be its spectral measure over* $\mathbb{R}^d$. *The* radial measure ν_k *of* k *is defined as the probability distribution of* $\|\omega\|_2$ *where* $\omega\sim\mu_\mathrm{k}$.

Thus, the radial measure can be realized by drawing a sample from the spectral measure and taking its Euclidean norm; however, this may not be the most efficient way to sample from the radial measure, as we discuss in Section A.6. If the kernel is radial, then the angular part of its spectral measure is uniform on the unit sphere, and thus the radial measure fully "encodes" the kernel.

The random map F_{k} in Theorem 1.2 is a close variant of the map from Section 3.2. Let $R_1, \ldots, R_m \sim \nu_k$ be iid samples from the radial measure of k. Using the notation from Section 3.2, we denote $\bar{V} = \|G\|_F^{-1} V = \|G\|_F^{-1} \sqrt{d} HGHB$. The map F_{k} is defined as

$$F_{\mathrm{k}}(x) = \frac{1}{\sqrt{m}} \oplus_{i=1}^{m} \left(\sin\left(R_i(\bar{V}x)_i\right), \cos\left(R_i(\bar{V}x)_i\right) \right).$$

Note that B and G take time $O(d)$ to sample, and $R_1, \ldots, R_m$ take time $O(m(1+T_\nu))$ to sample, and therefore the total time to sample F_{k} is $O(d + m(1+T_\nu))$.

4.1 Preliminaries

Throughout this section, we let $\mathcal{U}(d)$ denote the uniform distribution over the unit sphere $\mathbb{S}^{d-1}$. The following is the classical Schoenberg representation for positive definite radial kernels (Schoenberg, 1938a;b), which expresses them as mixtures of spherical shell kernels of different radii.

Lemma 4.2 (Schoenberg representation; see Section A.2)**.** *For every fixed dimension d, there exists a map $\Omega_d : \mathbb{R}_{\geq 0} \to [-1, 1]$ (called the spherical shell kernel) that satisfies*

$$\forall\, x, y \in \mathbb{R}^d, \quad \Omega_d(\|x-y\|_2) = \mathbb{E}_{\Theta \sim \mathcal{U}(d)}[\cos\left(\Theta^T(x-y)\right)].$$

If $\mathrm{k}(x,y) = \kappa(\|x-y\|_2)$ is a continuous positive definite normalized radial kernel, then

$$\forall\, x, y \in \mathbb{R}^d, \qquad \mathrm{k}(x,y) = \mathbb{E}_{R \sim \nu_k}[\Omega_d(R \cdot \|x-y\|_2)], \tag{19}$$

where ν_{k} is the radial measure of k*. Consequently,*

$$\forall\, x, y \in \mathbb{R}^d, \quad \mathrm{k}(x,y) = \mathbb{E}_{R \sim \nu_k,\, \Theta \sim \mathcal{U}(d)}[\cos(R \cdot \Theta^T(x-y))].$$

Lemma 4.3 (rank-1 spherical Brascamp-Lieb; see Section A.1)**.** *Under the same conditions as Lemma 3.6, for all non-negative measurable $f_i : \mathbb{R} \to [0, \infty)$,*

$$\mathbb{E}_{\Theta \sim \mathcal{U}(d)} \left[\prod_{i=1}^{k} f_i(w_i^T \Theta)^{c_i/2} \right] \leq \prod_{i=1}^{k} \left(\mathbb{E}_{\Theta \sim \mathcal{U}(d)}[f_i(w_i^T \Theta)] \right)^{c_i/2}.$$

4.2 Proof of Theorem 1.2

Fix a pair $x, y \in \mathbb{R}^d$ and $z = x - y$. Let $g = (g_1, \ldots, g_d) \sim N(0, I_d)$ be the diagonal of G. Let $\bar{g} := \|g\|_2^{-1} \cdot g$. By the spherical invariance of the normal distribution, $\bar{g}$ is distributed like $\mathcal{U}(d)$. We start by proving Equation (3). The following lemma follows from the same proof as in Lemma 3.7.

Lemma 4.4. $F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y) = \frac{1}{m} \sum_{i=1}^{m} \cos(R_i(\bar{V}z)_i)$.

The following lemma is analogous to Lemma 3.3.

Lemma 4.5. *Each row of $\bar{V}$ is marginally distributed like $\mathcal{U}(d)$.*

Proof. Fixing a row index i, let $h_i \in \{\pm 1\}^d$ be the ith column of $\sqrt{d}H$. Then the ith row of $\bar{V}$ can be written as $e_i^T \bar{V} = h_i^T \|G\|_F^{-1} GHB = \bar{g}^T \mathrm{diag}(h_i) HB$. Since H is an orthogonal matrix then so is $\mathrm{diag}(h_i)HB$. Therefore, since $\bar{g}$ is distributed like $\mathcal{U}(d)$, so is $\bar{g}^T \mathrm{diag}(h_i)HB$. □

By Lemma 4.4, $F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y)$ is the average of random variables of the form $\cos(R \cdot \Theta^T z)$, where R is sampled from ν_{k} and Θ from $\mathcal{U}(d)$. By Lemma 4.2, the mean of each such random variable is $\mathrm{k}(x,y)$, and Equation (3) follows.

We proceed to prove the concentration bound in Equation (4), whose proof at this point is rather similar to that of the Gaussian kernel case in Section 3.2. Assume $z \neq 0$ since otherwise Equation (4) holds trivially. Let $s, t > 0$. Conditioning on B and $\{R_i\}$ (so that the only remaining source of randomness in F_{k} is G), we have by Markov's inequality followed by Lemma 4.4 (analogously to Equation (13)),

$$\Pr\left[F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y) - \frac{1}{m} \sum_{i=1}^{m} \Omega_d(R_i \|z\|_2) \geq t \mid B, \{R_i\} \right]$$

$$= \Pr\left[\exp\left(s\left(F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y) - \frac{1}{m}\sum_{i=1}^{m}\Omega_d(R_i\|z\|_2)\right)\right) \geq \exp(st) \mid B, \{R_i\}\right]$$

$$\leq \exp(-st)\cdot\mathbb{E}\left[\exp\left(s\left(F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y) - \frac{1}{m}\sum_{i=1}^{m}\Omega_d(R_i\|z\|_2)\right)\right) \mid B, \{R_i\}\right]$$

$$= \exp(-st)\cdot\mathbb{E}\left[\prod_{i=1}^{m}\exp\left(\frac{s}{m}\left(\cos\left(R_i(\bar{V}z)_i\right) - \Omega_d\left(R_i\|z\|_2\right)\right)\right) \mid B, \{R_i\}\right]. \tag{20}$$

We now apply the spherical Brascamp-Lieb inequality (Lemma 4.3). We use the same w_i, c_i and c^* defined in Section 3.2. There we showed they satisfy the conditions of Lemma 3.6, which are the same as the conditions of Lemma 4.3. We define the maps f_i as

$$f_i(\alpha) := \exp\left(\frac{2}{c^*}\cdot\frac{s}{m}\left(\cos\left(R_i\|z\|_2\cdot\alpha\right) - \Omega_d\left(R_i\|z\|_2\right)\right)\right).$$

Recalling from Equation (14) that $(Vz)_i = \|z\|_2 w_i^T g$, and that $\bar{V} = \|G\|_F^{-1}V = \|g\|_2^{-1}V$ and $\bar{g} = \|g\|_2^{-1}g$, we have $w_i^T\bar{g} = \|z\|_2^{-1}(\bar{V}z)_i$. Consequently (recalling that $c_i = c^*$ for all i),

$$f_i(w_i^T\bar{g})^{c_i/2} = \exp\left(\frac{s}{m}\left(\cos\left(R_i(\bar{V}z)_i\right) - \Omega_d\left(R_i\|z\|_2\right)\right)\right).$$

Since $\bar{g}\sim\mathcal{U}(d)$, by applying Lemma 4.3 we get that the right-hand of Equation (20) is at most

$$\exp(-st)\cdot\prod_{i=1}^{m}\left(\mathbb{E}_{\Theta\sim\mathcal{U}(d)}[f_i(w_i^T\Theta)]\right)^{c^*/2}. \tag{21}$$

Next we control each $\mathbb{E}_{\Theta\sim\mathcal{U}(d)}[f_i(w_i^T\Theta)]$. Let X_i be the random variable $X_i = \cos(R_i(\bar{V}z)_i)$. By Lemma 4.5, the ith row of $\bar{V}$ is distributed like Θ, hence X_i is distributed like $\cos(R_i\Theta^T z)$. By Lemma 4.2, $\mathbb{E}[X_i] = \Omega_d(R_i\|z\|_2)$. Furthermore, X_i is clearly bounded in $[-1,1]$. Therefore, by Hoeffding's lemma (Lemma 3.5),

$$\mathbb{E}_\Theta[f_i(w_i^T\Theta)] = \mathbb{E}_\Theta\left[\exp\left(\frac{2s}{c^*m}\left(X_i - \mathbb{E}[X_i]\right)\right)\right] \leq \exp\left(\frac{2s^2}{(c^*)^2m^2}\right).$$

Plugging back this into Equations (21) and (20), and choosing $s := tc^*m/2$,

$$\Pr\left[F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y) - \frac{1}{m}\sum_{i=1}^{m}\Omega_d(R_i\|z\|_2) \geq t \mid B, \{R_i\}\right] \leq \exp\left(-st + \frac{s^2}{c^*m}\right) \leq \exp\left(-\frac{c^*mt^2}{4}\right).$$

The proof of the lower tail is similar, thus

$$\Pr\left[\left|F_{\mathrm{k}}(x)^T F_k k(y) - \frac{1}{m}\sum_{i=1}^{m}\Omega_d(R_i\|z\|_2)\right| \geq t \mid B, \{R_i\}\right] \leq 2\exp\left(-\frac{c^*mt^2}{4}\right).$$

Similarly to Section 3.2, by Fast JL flattening (Lemma 3.1) we have $c^* \geq 1/(2\log(2d/\eta))$ with probability $1-\eta$ over the choice of B. Therefore,

$$\Pr\left[\left|F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y) - \frac{1}{m}\sum_{i=1}^{m}\Omega_d(R_i\|z\|_2)\right| \geq t \mid \{R_i\}\right] \leq \eta + 2\exp\left(-\frac{mt^2}{8\log(2d/\eta)}\right).$$

Finally, we handle the term $\frac{1}{m}\sum_{i=1}^{m}\Omega_d(R_i\|z\|_2)$. Consider each $Y_i := \Omega_d(R_i\|z\|_2)$ as a push-forward random variable of R_i. By Lemma 4.2, Y_i is bounded in $[-1,1]$ and $\mathbb{E}_R[Y_i] = \mathrm{k}(x,y)$. Therefore, by the standard Chernoff-Hoeffding inequality,

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m}\Omega_d(R_i\|z\|_2) - \mathrm{k}(x,y)\right| \geq t\right] = \Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m}Y_i - \mathrm{k}(x,y)\right| \geq t\right] \leq 2\exp\left(-\frac{mt^2}{2}\right).$$

By the triangle inequality and a union bound,

$$\Pr\left[\left|F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y) - \mathrm{k}(x,y)\right| \geq 2t\right] \leq \eta + 2\exp\left(-\frac{mt^2}{8\log(2d/\eta)}\right) + 2\exp\left(-\frac{mt^2}{2}\right).$$

Scaling t to $t/2$ yields Equation (4), completing the proof of Theorem 1.2.

AI USE STATEMENT

All the proofs in this paper were found through interactions with generative AI, primarily GPT 5.6 Sol, and rely heavily on its contribution along with substantial human input. The authors wrote the manuscript manually based on those interactions to ensure correctness, readability and faithfulness to existing literature. They are accountable for the content and correctness.

REFERENCES

Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

Noga Alon. Perturbed identity matrices have high rank: Proof and applications. *Combinatorics, Probability and Computing*, 18(1-2):3–15, 2009.

Alex Andoni and Piotr Indyk. Dimension reduction in kernel spaces from locality-sensitive hashing. 2009.

Haim Avron, Vikas Sindhwani, Jiyan Yang, and Michael W Mahoney. Quasi-monte carlo feature maps for shift-invariant kernels. *Journal of Machine Learning Research*, 17(120):1–38, 2016.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of machine learning research*, 18(21):1–38, 2017.

Arturs Backurs, Piotr Indyk, and Tal Wagner. Space and time efficient kernel density estimation in high dimensions. *Advances in neural information processing systems*, 32, 2019.

Arturs Backurs, Zinan Lin, Sepideh Mahabadi, Sandeep Silwal, and Jakub Tarnawski. Efficiently computing similarities to private datasets. In *International Conference on Learning Representations (ICLR)*, 2024.

Yair Bartal, Ben Recht, and Leonard J Schulman. Dimensionality reduction: beyond the johnson-lindenstrauss bound. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pp. 868–887. SIAM, 2011.

Franck Barthe, Dario Cordero-Erausquin, Michel Ledoux, and Bernard Maurey. Correlation and Brascamp-Lieb inequalities for Markov semigroups. *arXiv preprint arXiv:0907.2858*, 2009.

Franck Barthe, Dario Cordero-Erausquin, Michel Ledoux, and Bernard Maurey. Correlation and Brascamp–Lieb inequalities for Markov semigroups. *International Mathematics Research Notices*, 2011(10):2177–2216, 2011. doi: 10.1093/imrn/rnq151.

Herm Jan Brascamp and Elliott H Lieb. Best constants in young's inequality, its converse, and its generalization to more than three functions. *Advances in Mathematics*, 20(2):151–173, 1976.

Moses Charikar and Paris Siminelakis. Hashing-based-estimators for kernel density in high dimensions. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science*, pp. 1032–1043. IEEE, 2017.

Moses Charikar, Michael Kapralov, Navid Nouri, and Paris Siminelakis. Kernel density estimation through density constrained near neighbor search. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*, pp. 172–183. IEEE, 2020.

Di Chen and Jeff M Phillips. Relative error embeddings of the gaussian kernel distance. In *International Conference on Algorithmic Learning Theory*, pp. 560–576. PMLR, 2017.

Wei-Kuo Chen, Nikos Dafnis, and Grigoris Paouris. Improved hölder and reverse hölder inequalities for gaussian random vectors. *Advances in Mathematics*, 280:643–689, 2015.

Yeshwanth Cherapanamjeri and Jelani Nelson. Uniform approximations for randomized hadamard transforms with applications. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 659–671, 2022.

Krzysztof Choromanski and Vikas Sindhwani. Recycling randomness with structure for sublinear time kernel expansions. In *International Conference on Machine Learning*, pp. 2502–2510. PMLR, 2016.

Krzysztof M Choromanski, Mark Rowland, and Adrian Weller. The unreasonable effectiveness of structured random orthogonal embeddings. *Advances in neural information processing systems*, 30, 2017.

DLMF. NIST Digital Library of Mathematical Functions. https://dlmf.nist.gov/. National Institute of Standards and Technology.

Chang Feng, Qinghua Hu, and Shizhong Liao. Random feature mapping with signed circulant matrix projection. In *IJCAI*, pp. 3490–3496, 2015.

Alex Gittens and Michael W Mahoney. Revisiting the nyström method for improved large-scale machine learning. *The Journal of Machine Learning Research*, 17(1):3977–4041, 2016.

L Golinskii, M Malamud, and L Oridoroga. Radial positive definite functions and schoenberg matrices with negative eigenvalues. *Transactions of the American Mathematical Society*, 370(1):1–25, 2018.

Zhen Huang, Jiajin Sun, and Yian Huang. Quasi-monte carlo features for kernel approximation. In *Forty-first International Conference on Machine Learning*, 2024.

Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the international conference on machine learning*, 2013.

Elliott H Lieb. Gaussian kernels have only gaussian maximizers. *Inventiones mathematicae*, 102 (1):179–208, 1990.

Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan AK Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, 2021.

George Marsaglia and Wai Wan Tsang. A simple method for generating gamma variables. *ACM Transactions on Mathematical Software (TOMS)*, 26(3):363–372, 2000.

Marina Munkhoeva, Yermek Kapushev, Evgeny Burnaev, and Ivan Oseledets. Quadrature-based features for kernel approximation. *Advances in neural information processing systems*, 31, 2018.

Duke Nguyen, Du Yin, Aditya Joshi, and Flora Salim. Spectraformer: A unified random feature framework for transformer. *ACM Transactions on Intelligent Systems and Technology*, 17(3): 1–29, 2026.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah A Smith, and Lingpeng Kong. Random feature attention. *arXiv preprint arXiv:2103.02143*, 2021.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in neural information processing systems*, 28, 2015.

Issei Saito, Masatoshi Nagano, Tomoaki Nakamura, Daichi Mochihashi, and Koki Mimura. Scalable unsupervised segmentation via random fourier feature-based gaussian process. In *IECON 2025–51st Annual Conference of the IEEE Industrial Electronics Society*, pp. 1–7. IEEE, 2025.

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*, pp. 143–152. IEEE, 2006.

Isaac J Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39 (4):811–841, 1938a.

Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938b.

Paris Siminelakis, Kexin Rong, Peter Bailis, Moses Charikar, and Philip Levis. Rehashing kernel evaluation in high dimensions. In *International Conference on Machine Learning*, pp. 5789–5798, 2019.

Tal Wagner. New bounds for kernel sums via fast spherical embeddings. In *International Conference on Machine Learning (ICML)*. PMLR, 2026.

Felix Xinnan X Yu, Ananda Theertha Suresh, Krzysztof M Choromanski, Daniel N Holtmann-Rice, and Sanjiv Kumar. Orthogonal random features. *Advances in neural information processing systems*, 29, 2016.

Haoran Zhang and Feng Zhou. Flexformer: Flexible linear transformer with learnable attention kernel. *arXiv preprint arXiv:2606.27748*, 2026.

A Omitted proofs and details

A.1 Brascamp-Lieb inequalities

In this appendix, for completeness, we give more precise references for the Brascamp-Lieb inequalities in Lemmas 4.3 and 3.6.

Gaussian Brascamp-Lieb (Lemma 3.6). This lemma is a slight adaptation of a statement from Barthe et al. (2011), or Theorem 1(i) from Chen et al. (2015). We show how to derive it from the statement in the bottom of page 5 in the arxiv version Barthe et al. (2009) of Barthe et al. (2011). If one takes $n_i = 1$ for all i, meaning each E_i is a one-dimensional subspace spanned by the unit vector w_i, then that statement becomes exactly our lemma except that it requires the equality

$$\sum_{i=1}^{k} c_i w_i w_i^T = I_d$$

instead of the relaxed condition in Equation (10). Showing that the relaxed condition indeed suffices is straightforward: suppose we have Equation (10). Denote the remainder

$$R = I_d - \sum_{i=1}^{k} c_i w_i w_i^T,$$

and write its spectral decomposition

$$R = \sum_{i=k+1}^{k'} c_i w_i w_i^T,$$

where $c_i > 0$ and w_i are unit vectors. Now we have the equality condition

$$\sum_{i=1}^{k'} c_i w_i w_i^T = I_d.$$

For every $i = k+1, \ldots, k'$ we choose $f_i \equiv 1$, which obviously doesn't change the conclusion of the lemma.

Spherical Brascamp-Lieb (Lemma 4.3). This lemma is exactly the rank-1 case of the *"Geometric Brascamp-Lieb inequality on the sphere"* in Barthe et al. (2009), page 6; that is, where each E_i (in their notation) is the 1-dimensional subspace spanned by w_i (in our notation).

A.2 SCHOENBERG'S REPRESENTATION

Lemma 4.2 is classical and appears in various phrasings in the literature. We briefly point out how to obtain it from Theorem 1.2 in Golinskii et al. (2018). Their equation (1.2) in our notation is $\kappa(r) = \mathbb{E}_R[\Omega_d(rR)]$ for all $r \geq 0$, and by their equation (1.4), $\Omega_d(\|z\|_2) = \mathbb{E}_\Theta[e^{\mathbf{i}\cdot\Theta^T z}]$ for all $z \in \mathbb{R}^d$. Since $e^{\mathbf{i}\zeta} = \cos(\zeta) + \mathbf{i}\sin(\zeta)$ and sin is symmetric about zero, the latter is equivalent to $\Omega_d(\|z\|_2) = \mathbb{E}_\Theta[\cos(\Theta^T z)]$. Hence also $\Omega_d(R\cdot\|z\|_2) = \mathbb{E}_\Theta[\cos(R\cdot\Theta^T z)]$ for all $R \geq 0$. Plugging $z = x - y$ and $r = \|z\|_2$ yields the lemma.

A.3 OTHER OMITTED PROOFS

Proof of Lemma 3.2. Since H is orthogonal and B only flips some signs of z, $\|HBz\|_2 = \|Bz\|_2 = \|z\|_2$. □

Proof of Lemma 3.3. Conditioned on B, the vector $u = HBz$ is fixed and all randomness in V is encompassed in G. Since G has iid gaussians on its diagonal, and $\sqrt{d}H$ has all entries in ± 1, each row of $\sqrt{d}HG$ is a vector of iid gaussians (while the different rows are not independent). Therefore, by the spherical invariance of the normal distribution, each entry of $Vz = \sqrt{d}HGu$ is distributed like $N(0, \|u\|_2^2)$, which per Lemma 3.2 is $N(0, \|z\|_2^2)$. □

Proof of Lemma 3.7.

$$\begin{aligned} F(x)^T F(y) &= \frac{1}{m}\sum_{i=1}^{m} \left(\sin((Vx)_i)\sin((Vy)_i) + \cos((Vx)_i)\cos((Vy)_i)\right) \\ &= \frac{1}{m}\sum_{i=1}^{m} \left(\cos((Vx)_i - (Vy)_i)\right) \\ &= \frac{1}{m}\sum_{i=1}^{m} \cos((Vz)_i), \end{aligned}$$

by a standard trigonometric sum-product formula for the middle equality. □

A.4 THE $m > d$ REGIME

Section 3.2 proves Theorem 1.1 in the case $m < d$. In this appendix we show that the extension for $m > d$ follows readily.

Suppose $m > d$, and write $m = \sum_{\ell=1}^{q} m_j$ with $q = \lceil m/d \rceil$ and $1 \leq m_j \leq d$ for all j. Sample the map $F^* : \mathbb{R}^d \to \mathbb{R}^{2m}$ as follows:

- Sample a random diagonal matrix $B \in \mathbb{R}^{d\times d}$ with iid random signs on the diagonal.
- For every $j = 1, \ldots, q$, sample a random diagonal matrix $G_j \in \mathbb{R}^{d\times d}$ with iid standard Gaussians on the diagonal. The $\{G_j\}$ are independent.
- Denote $V_j = \sqrt{d}HG_jH_B$. Define $F_j : \mathbb{R}^d \to \mathbb{R}^{2m_j}$ as the map from Section 3.2:

$$F_j(x) = \frac{1}{\sqrt{m_j}} \oplus_{i=1}^{m_j} \left(\sin((V_j x)_i),\ \cos((V_j x)_i)\right).$$

F^* is now defined as

$$F^* = \oplus_{i=j}^{q} \sqrt{\frac{m_j}{m}} F_j.$$

Fix a pair $x, y \in \mathbb{R}^d$ and let $z = x - y$. For brevity, denote

$$\Delta_j = \frac{1}{m_j}\sum_{i=1}^{m_j} \left(\cos((V_j z)_i) - \exp(-\|z\|_2^2/2)\right).$$

Then by the MGF bound in Equation (17),

$$\mathbb{E}[\exp(s\Delta_j) \mid B] \leq \exp\left(\frac{s^2}{2c^* m_j}\right)$$

Let

$$\Delta^* = \sum_{j=1}^{q} \frac{m_j}{m} \Delta_j.$$

Similarly to Lemma 3.7, one can show that

$$F^*(x)^T F^*(y) = \frac{1}{m} \sum_{j=1}^{q} \sum_{i=1}^{m_j} \cos((V_j z)_i),$$

and therefore

$$\Delta^* = F^*(x)^T F^*(y) - \exp(-\|z\|_2^2/2). \tag{22}$$

Observe that $\{V_j\}$ are independent conditioned on B, and also that c^* is mutual to all $\{F_j\}$, since it depends only on B and z. Therefore,

$$\begin{aligned}
\mathbb{E}[\exp(s\Delta^*) \mid B] &= \mathbb{E}\left[\exp\left(s \sum_{j=1}^{q} \frac{m_j}{m} \Delta_j\right) \mid B\right] \\
&= \prod_{j=1}^{q} \mathbb{E}\left[\exp\left(\frac{s m_j}{m} \Delta_j\right) \mid B\right] \\
&\leq \prod_{j=1}^{q} \exp\left(\frac{(s m_j/m)^2}{2c^* m_j}\right) \\
&= \exp\left(\frac{s^2}{2c^* m^2} \sum_{j=1}^{q} m_j\right) \\
&= \exp\left(\frac{s^2}{2c^* m}\right).
\end{aligned}$$

Plugging this MGF bound in the Hoeffding-Chernoff blueprint and using Equation (22), we get that for every $s, t > 0$,

$$\Pr\left[F^*(x)^T F^*(y) - \exp(-\|z\|_2^2/2) > t \mid B\right] \leq \exp\left(-st + \frac{s^2}{2c^* m}\right).$$

From this point, the proof is finished in the same way as Section 3.2: we choose $s = mc^* t$ and de-condition on B with the JL flattening event from Lemma 3.1. This proves the upper tail bound. The proof of the lower tail is similar. The extension of Theorem 1.2 to $m \geq d$ follows from similar arguments.

A.5 Proof of Theorem 1.3

We will show a lower on the output dimension m in Theorem 1.3.

Lemma A.1. *Under the assumptions of Theorem 1.3,* $m = \Omega(\frac{\log(1/\delta))}{\varepsilon^2 \log(1/\varepsilon)})$.

Theorem 1.3 follows from Lemma A.1 because computing F requires at least reading its d input coordinates (this argument is very simple and omitted here) and writing its output m coordinates, thus overall $\Omega(d + m)$ time.

The proof is Lemma A.1 is by the following classical theorem on the rank of perturbed identity matrices.

Theorem A.2 (Alon (2009))**.** *Let* $M \in \mathbb{R}^{n \times n}$ *such that* $|M_{ii}| \geq \frac{1}{2}$ *for all* i *and* $|M_{ij}| \leq \alpha$ *for all* $i \neq j$. *Let* $\alpha \in [\frac{1}{2\sqrt{n}}, \frac{1}{4})$. *Then* $\operatorname{rank}(M) \geq \Omega(\frac{\log n}{\alpha^2 \log(1/\alpha)})$.

Proof of Lemma A.1. Let $n = \lceil 1/2\sqrt{\delta} \rceil$. Let $x_i = i\sqrt{2\log(2/\varepsilon)}\, e_1$ for $i \in [n]$, where e_1 is the first standard basis vector in $\mathbb{R}^d$. The pairwise Gaussian kernel values given by $K_{ij} = \exp(-\|x_i - x_j\|_2^2/2)$ satisfy

$$\forall i,\ K_{ii} = 1 \quad \text{and} \quad \forall i \neq j,\ 0 \leq K_{ij} \leq \varepsilon/2. \tag{23}$$

Since $n^2\delta \leq 1/2$, a union bound gives the existence of a function F whose Gram matrix $G_{ij} = F(x_i)^T F(x_j)$ satisfies, by combining Equations (5) and (23)

$$\forall i,\ G_{ii} \geq 1 - \varepsilon \geq \frac{1}{2} \quad \text{and} \quad \forall i \neq j,\ |G_{ij}| \leq 3\varepsilon/2.$$

Furthermore, $\operatorname{rank}(G) \leq m$. Choosing $\alpha = 3\varepsilon/2$, the assumptions of Theorem 1.3 guarantee that

$$\frac{1}{2\sqrt{n}} \leq \delta^{1/4} \leq \varepsilon \leq \alpha \leq \frac{1}{4}.$$

Therefore, by Theorem A.2,

$$m \geq \operatorname{rank}(G) \geq \Omega\left(\frac{\log(1/\delta)}{\varepsilon^2 \log(1/\varepsilon)}\right),$$

as desired. □

A.6 Applications to Kernels Beyond Gaussian

We give explicit samplers for the radial measures of several standard kernels, yielding efficient implementations and corollaries of Theorem 1.2. The standard kernels we consider are the exponential, rational quadratic, and Matérn kernels. For each of these kernels, sampling from the radial measure requires sampling only two independent scalar Gamma random variables. We assume sampling such Gamma random variables take $O(1)$ expected time as in Marsaglia & Tsang (2000). Consequently, Theorem 1.2 gives randomized feature maps for these kernels with expected sampling time $O(d+m)$ and evaluation time $O((d+m)\log d)$, while retaining the unbiasedness and concentration guarantees of the theorem.

We write $\mathrm{k}(x, y) = \kappa(r)$, where $r = \|x - y\|_2$. In the following examples, we can set the bandwidth equal to 1 without loss of generality. For $a > 0$, let $\mathrm{Gamma}(a, 1)$ denote the Gamma distribution with shape parameter a and rate parameter 1.

Recall from Definition 4.1 that the radial measure ν_{k} is the distribution of $\|\omega\|_2$ when $\omega \sim \mu_{\mathrm{k}}$, where μ_{k} is the spectral measure of k. We describe how to generate a scalar radius $R \sim \nu_{\mathrm{k}}$, which supplies each of the independent radii $R_1, \ldots, R_m$ used in the construction of the feature map F_{k} of Theorem 1.2.

Our samplers follow from the following observation. If a nonnegative random variable S satisfies

$$\kappa(r) = \mathbb{E}_S\left[\exp(-Sr^2/2)\right] \qquad \text{for all } r \geq 0,$$

then drawing $Q \sim \mathrm{Gamma}(d/2, 1)$ independently of S and setting $R = \sqrt{2SQ}$ gives a sample from ν_{k}. Indeed, for $Z \sim N(0, I_d)$ independent of S, the vector $\sqrt{S}Z$ has distribution μ_{k}, and $\|Z\|_2^2/2$ has distribution $\mathrm{Gamma}(d/2, 1)$. In the constructions below, we generate S from a scalar Gamma variable A, whose parameters depend on the kernel and which is sampled independently of Q.

Exponential kernel. The exponential kernel is defined as $\mathrm{k}(x, y) = \exp(-\|x - y\|_2)$. For $A \sim \mathrm{Gamma}(1/2, 1)$, integrating its density gives

$$\mathbb{E}_A\left[\exp\left(-\frac{r^2}{4A}\right)\right] = \frac{1}{\sqrt{\pi}} \int_0^\infty a^{-1/2} \exp\left(-a - \frac{r^2}{4a}\right) da = \exp(-r),$$

where the final equality follows from standard calculations. Thus we take $S = 1/(2A)$, and a sample from the radial measure is $R = \sqrt{Q/A}$.

Rational quadratic kernel. For a shape parameter $\alpha > 0$, the rational quadratic kernel is

$$\mathrm{k}(x, y) = \left(1 + \frac{\|x - y\|_2^2}{2\alpha}\right)^{-\alpha}.$$

Taking $A \sim \mathrm{Gamma}(\alpha, 1)$, its Laplace transform gives $\mathbb{E}_A[\exp(-Ar^2/(2\alpha))] = (1+r^2/(2\alpha))^{-\alpha}$. Hence $S = A/\alpha$, and a sample from the radial measure is $R = \sqrt{2QA/\alpha}$. The case $\alpha = 1$ is the Cauchy kernel $\mathrm{k}(x, y) = (1 + \|x - y\|_2^2/2)^{-1}$ in this parameterization.

Matérn kernel. For a smoothness parameter $\beta > 0$, the Matérn kernel is defined by

$$\kappa(r) = \frac{2^{1-\beta}}{\Gamma(\beta)} \left(\sqrt{2\beta}\, r\right)^{\beta} K_\beta \left(\sqrt{2\beta}\, r\right), \qquad r > 0,$$

with $\kappa(0) = 1$ by continuity. Here Γ is the Gamma function and K_β is the modified Bessel function of the second kind. The integral representation of K_β (see DLMF) gives $\kappa(r) = \mathbb{E}_A[\exp(-\beta r^2/(2A))]$ for $A \sim \mathrm{Gamma}(\beta, 1)$. Thus $S = \beta/A$, and a sample from the radial measure is $R = \sqrt{2\beta Q/A}$. The exponential kernel is the special case $\beta = 1/2$.

Each sampler uses two independent scalar Gamma samples. Using the standard Gamma sampler of Marsaglia & Tsang (2000), each radius takes expected $O(1)$ time under the usual model of unit-cost real arithmetic, elementary functions, and scalar uniform and Gaussian sampling. For every R_i, we draw a fresh independent pair (Q, A), independently of B and G, the random sign and Gaussian diagonal matrices defined in Section 3.2.

Consequently, the radial sampling cost $T_{\mathrm{k}}^{\mathrm{rad}}$ in Theorem 1.2 is expected $O(1)$ for each of these three kernels. The theorem therefore gives a randomized map $F_{\mathrm{k}} : \mathbb{R}^d \to \mathbb{R}^{2m}$ with expected sampling time $O(d + m)$ and evaluation time $O((d + m) \log d)$. For every fixed pair $x, y \in \mathbb{R}^d$, the estimator $F_{\mathrm{k}}(x)^T F_{\mathrm{k}}(y)$ is unbiased for $\mathrm{k}(x, y)$ and satisfies the concentration bound of Theorem 1.2.

A.7 Kernel density estimation

Let $\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel. Given a subset $X \subset \mathbb{R}^d$ of size $n > 0$, the kernel density estimation (KDE) function is defined as $KDE_X(y) = \frac{1}{n} \sum_{x \in X} \mathrm{k}(x, y)$. A KDE data structure has the following operations:

- Initialize: given $X \subset \mathbb{R}^d$, initialize the data structure.
- Query: given $y \in \mathbb{R}^d$, return an approximation $\widetilde{K}_X(y)$ of $KDE_X(y)$.

The operations may randomized. We say that the data structure satisfies (ε, δ)-additive approximation if after initialization with any X it satisfies

$$\forall\, y \in \mathbb{R}^d, \quad \Pr\left[\left|KDE_X(y) - \widetilde{K}_X(y)\right| > \varepsilon\right] \le \delta,$$

where the probability is over the internal randomness of the KDE data structure.

The following result shows the existence of a Gaussian KDE data structure that satisfies (ε, δ)-additive approximation with query time $\tilde{O}(d + \varepsilon^{-2})$. It is a corollary of Theorem 1.1. It answers the main open question posed in Wagner (2026); see that reference there for an account of previous bounds for this problem. We note that there is also a distinct line of work for KDE data structures with multiplicative error. The algorithms and bounds with this type of error are different than the pure additive error setting of this section, and we refer to (Charikar & Siminelakis, 2017; Backurs et al., 2019; Charikar et al., 2020) for details about this different line of approach.

Theorem A.3. *For all $\varepsilon, \delta \in (0, 1/2)$, there is a KDE data structure for the Gaussian kernel with (ε, δ)-additive error, with the following computational bounds:*

- *initialization time: $O(|X|(d + m) \log d)$,*
- *memory size: $O(d + m)$,*
- *query time: $O((d + m) \log d)$,*

where $m = O(\varepsilon^{-2}\log(d/\varepsilon\delta)\log(d/\varepsilon\delta))$.

Proof. Assume w.l.o.g. that d is a power of 2. Let $m \geq 16\varepsilon^{-2}\log(16d/\varepsilon\delta)\log(16/\varepsilon\delta))$. To initialize the data structure, we draw a randomized map $F : \mathbb{R}^d \to \mathbb{R}^{2m}$ from Theorem 1.1 and compute $F(X) = \frac{1}{|X|}\sum_{x\in X} F(x)$. This takes time $O(|X|(d+m)\log d)$ by the sampling and evaluation times of F in Theorem 1.1. We store the map F (this takes $O(d+m)$ memory) and the value $F(x)$ (this takes $O(m)$ memory). At query time, given $y \in \mathbb{R}^d$, we compute $F(y)$ (time $O((d+m)\log d)$) and return $\widetilde{K}_X(y) = F(X)^T F(y)$ (time $O(m)$).

To show the (ε, δ)-additive approximation guarantee, we choose $p = \varepsilon\delta/4$ and $\eta = p/2$. Then by Equation (2),

$$\forall\, x \in X, \quad \Pr\left[\left|F(x)^T F(y) - \exp(-\|x-y\|_2^2/2)\right| > \varepsilon/2\right] \leq p. \tag{24}$$

Denote,

$$E(x) = \left|F(x)^T F(y) - \exp(-\|x-y\|_2^2/2)\right| \quad \text{and} \quad Z = \frac{1}{|X|}\sum_{x\in X} \mathbf{1}\{E(x) > \varepsilon/2\}.$$

Then $\mathbb{E}[Z] \leq p$ by Equation (24), and hence by Markov's inequality, $\Pr[Z > \varepsilon/4] \leq 4p/\varepsilon = \delta$. Next, observe that $\|F(x)\|_2^2 = 1$ for all $x \in \mathbb{R}^d$ (since $F(x)$ has the form $\frac{1}{\sqrt{m}} \oplus_{j=1}^{m} (\sin(\theta_j), \cos(\theta_j))$ for some $\{\theta_j\}$), and hence $|F(x)^T F(y)| \leq 1$ for all x, y by Cauchy-Schwartz. Furthermore, $\exp(-\|x-y\|_2^2/2) \in [0,1]$ for all x, y. Consequently, $|E(x)| \leq 2$ for every $x \in X$. Therefore, for every fixed y,

$$\begin{aligned}
\left|\widetilde{K}_X(y) - KDE_X(y)\right| &= \left|\frac{1}{|X|}\sum_{x\in X}\left(F(x)^T F(y) - \exp(-\|x-y\|_2^2/2)\right)\right| \\
&\leq \frac{1}{|X|}\sum_{x\in X} E(x) \\
&= \frac{1}{|X|}\sum_{x\in X:\ E(x)\leq\varepsilon/2} E(x) + \frac{1}{|X|}\sum_{x\in X:\ E(x)>\varepsilon/2} E(x) \\
&\leq \frac{\varepsilon}{2} + 2Z.
\end{aligned}$$

Since $\Pr[Z > \varepsilon/4] \leq \delta$, the theorem follows. □

A similar theorem follows for KDE with other radial kernels by Theorem 1.2, with the corresponding radial-sampling cost during initialization. The query time and memory usage remain unchanged.