

Diameter-Free Pointwise Gaussian KDE in Near-Additive Query Time*

Matěj Kripner Václav Rozhoň Robert Šámal
Martin Šindelář Adrián Zámečník

Abstract

Wagner asked whether Gaussian kernel density estimation in $\mathbb{R}^d$ admits query time $\tilde{O}(d + \varepsilon^{-2})$ without dependence on the diameter of the data. We answer this affirmatively for each fixed query: a randomized data structure answers a query to additive error ε with probability $1 - \delta$ in time $O((d + \varepsilon^{-2}) \log(d + 1) \log(2/\delta))$, independently of the number of data points and of their diameter. The key ingredient is a variance bound for a Fastfood-type structured Fourier block that holds uniformly in the norm of the displacement.

1 Introduction

1.1 Background and the source problem

For a nonempty finite dataset $X \subset \mathbb{R}^d$ and bandwidth $\sigma > 0$, the Gaussian kernel mean at a query y is

$$K_X(y) = \frac{1}{|X|} \sum_{x \in X} k_\sigma(x, y), \quad k_\sigma(x, y) = \exp(-\|x - y\|_2^2 / \sigma^2).$$

The computational problem is to preprocess X so that this mean can be approximated without scanning the dataset at each query. Dense random Fourier features [3] separate the dependence on $|X|$ from the query: store the average feature vector of the data, then take its inner product with the feature vector of y . For constant success probability, $O(\varepsilon^{-2})$ independent bounded features suffice for additive error ε . Evaluating their dense projections, however, costs $O(d\varepsilon^{-2})$. The desired improvement is to replace this product by a sum, up to logarithmic factors.

Wagner [1, Def. 1.1] formulates the guarantee for each fixed query, with probability over the randomized preprocessing. If data and queries lie in a region of diameter Δ , write $\Delta_\sigma = \Delta/\sigma$. His Table 1 compares the classical RFF bound with $\tilde{O}(d + \varepsilon^{-4})$ from fast dimension reduction followed by RFF [2, Thm. E.2 and Cor. E.3], and with $\tilde{O}(d + \Delta_\sigma^2 \varepsilon^{-2})$ from Fastfood-type features [8]. Wagner’s fast spherical embedding gives $\tilde{O}(d + \varepsilon \Delta_\sigma^2 + \varepsilon^{-3})$ [1, Thms. 1.2–1.3]. The discussion then asks the following question. The section number in its heading refers to the arXiv preprint; the same question appears in Section 6 of the published ICML version.

Question 1.1 (Wagner, Section 7). Is a Gaussian KDE query-time bound $\tilde{O}(d + \varepsilon^{-2})$ possible with only polylogarithmic, or no, dependence on the effective diameter Δ_σ ? More generally, what is the optimal trade-off between ε and Δ_σ ?

Theorem 2.1 gives an affirmative answer to the existence question with no diameter dependence. It uses the pointwise quantifiers of Wagner’s Definition 1.1: y is fixed independently of the preprocessing randomness. The theorem is not an optimality lower bound for all possible data structures, and it does not settle the stronger problem of simultaneous accuracy throughout a query domain.

*This work was produced within the Bolzano project (<https://bolzano.app>), using AI systems for mathematical research and manuscript preparation. The listed authors contributed to the project and the preparation of this release. The manuscript received only limited human checking.

1.2 Main results

The main statements are as follows.

- (1) Theorem 2.1 gives query time $O((d+\varepsilon^{-2})\log(d+1)\log(2/\delta))$ and storage $O((d+\varepsilon^{-2})\log(2/\delta))$, independently of $|X|$ and diameter. Preprocessing costs $|X|$ times the feature-evaluation bound.
- (2) Proposition 3.6 bounds the variance of one structured Fourier block by $5/D$ for every fixed displacement, regardless of its norm. Lemma 4.1 transfers this estimate to an arbitrary dataset.
- (3) Proposition 5.1 gives a one-block witness explaining why the sign-cancellation argument cannot be made uniform by choosing a displacement after the signs are revealed.

The computation model charges constant cost for real arithmetic and trigonometric evaluation, as specified in Section 2. No finite-precision bit-complexity or empirical speedup claim is made.

1.3 Why the diameter disappears

The fast map shares randomness among many features. Let D be the next power of two at least d , let H be the normalized Walsh–Hadamard matrix, and let G and B be independent diagonal matrices with independent standard Gaussian and Rademacher entries, respectively. After padding a displacement to $z \in \mathbb{R}^D$, use

$$W = \sqrt{D} H G H B, \quad Y(z) = \frac{1}{D} \sum_{i=1}^D \cos((Wz)_i).$$

Two fast Hadamard transforms evaluate an entire block in $O(D \log(D+1))$ time. Each row is marginally standard Gaussian, so $Y(z)$ is unbiased for $e^{-\|z\|_2^2/2}$. The difficulty is that these rows are correlated: their number alone does not guarantee a variance reduction by D .

The useful cancellation occurs before taking a supremum over displacements. Index coordinates by $\mathbb{F}_2^{\log_2 D}$. Conditional on B , rows separated by ℓ have projection covariance

$$\rho_\ell = \sum_r b_r b_{r+\ell} z_r z_{r+\ell}.$$

Their cosine covariance is $e^{-\|z\|_2^2}(\cosh \rho_\ell - 1)$. Bounding its growth in isolation loses the decay of the Gaussian kernel at large distances. Keeping that decay, Lemma 3.5 gives

$$e^{-\|z\|_2^2}(\cosh \rho_\ell - 1) \leq \frac{2\rho_\ell^2}{(1 + \|z\|_2^2)^2}.$$

The exact identity $\mathbb{E}_B \sum_{\ell \neq 0} \rho_\ell^2 = 2(\|z\|_2^4 - \|z\|_4^4)$ now cancels the potentially large norm factor. The resulting $5/D$ bound holds for every fixed z , with a constant independent of its length. Minkowski’s inequality lifts it to the data average even though all data points use the same map. Independent blocks give constant success probability, and independent median copies reduce the failure probability to δ .

1.4 Related work and the scope of the improvement

Fast kernel summation. In low dimension, the fast Gauss transform of Greengard and Strain [4] evaluates Gaussian sums by series expansions, with cost growing exponentially in d . In high dimension, hashing-based estimators of Charikar and Siminelakis [5] and their refinements [6] give multiplicative approximations with query time sublinear in $|X|$ that depends on a lower bound on the density. The present problem asks instead for additive error with query time independent of $|X|$, the regime of random Fourier features [3]. Fast structured projections built from randomized Hadamard transforms originate in the fast Johnson–Lindenstrauss transform of Ailon and Chazelle [7].

Fast structured Fourier features. The matrix above is a permutation-free Fastfood block without the additional radial scaling. Fastfood already supplies marginal Gaussianity, unbiasedness, fast evaluation and the conditional cosh covariance calculation [8, Lem. 3 and Thms. 4–5]. For its usual block, which also contains a random permutation, the published variance bound has a remainder $C(\alpha)/n$ for n frequencies at normalized distance α , where

$$C(\alpha) = 6\alpha^4(e^{-\alpha^2} + \alpha^2/3).$$

Its growth as $\alpha \rightarrow \infty$ does not give the diameter-free conclusion. The expanded Fastfood manuscript retains this remainder and also gives a concentration bound proportional to the displacement length [11, Thm. 9, Cor. 10, and Thm. 11]. The contribution here is the norm-uniform completion using damping and the exact signed autocorrelation identity, rather than a new general Fourier-feature architecture. The broader structured-matrix framework of Choromanski and Sindhvani includes Fastfood and circulant or Toeplitz-like maps; its Theorems 4.1–4.2 are stated for bounded inputs and involve coherence and structural parameters [10].

Orthogonal features. Orthogonal random features reduce Gaussian-kernel estimation error through dependence of a different kind [13]. Their scaled Haar construction and its variance analysis should be distinguished from the fast triple-Hadamard construction SORF. For SORF, Theorem 3 of [13] bounds bias by $6\alpha/\sqrt{d}$ and the accompanying discussion leaves tight variance and concentration guarantees open. Choromanski, Rowland, and Weller analyze structured orthogonal embeddings for dimensionality reduction and the angular kernel [14]. Further RBF variance comparisons include a Gaussian improvement for each fixed displacement length in sufficiently high dimension [15, Cor. 3.4]. These results explain important benefits of orthogonality but do not supply the all-norm variance estimate for the nonorthogonal Gaussian block used here. Radial scaling also matters: Demni and Kadri compute a Bessel-kernel expectation for Haar orthogonal features without Gaussian radial scaling [12, Thm. 2]; that model is distinct from the unbiased Gaussian features used here.

A signed-circulant variance claim. Feng, Hu, and Liao state an independent-feature variance formula for signed circulant projections [9, Thm. 2, Eq. (12)]. The following explicit calculation shows why that formula cannot substitute for the variance proof below. Let $d \geq 2$, let C be a circulant matrix generated by independent $N(0, \tau^2)$ entries, set $P = \text{diag}(s_1, \dots, s_d)C$, and take $v = (t/\sqrt{d})\mathbf{1}$ with $t, \tau > 0$. All coordinates of Cv equal one Gaussian variable Z of variance $\nu^2 = t^2\tau^2$. Since cosine is even,

$$\frac{1}{d} \sum_{j=1}^d \cos((Pv)_j) = \cos Z, \quad \text{Var}(\cos Z) = \frac{1}{2}(1 - e^{-\nu^2})^2.$$

There is no factor $1/d$. The zero covariance of the signed projections does not make them independent: their joint distribution is a mixture of Gaussian pairs. This counterexample concerns exactly the cosine estimator in Theorem 2; the random-phase feature map in that paper’s Eq. (9) is a separately defined estimator.

Coresets and query quantifiers. Uniform approximation by a small data subset is a different objective. Phillips and Tai obtain $O(\varepsilon^{-2})$ -size KDE coresets independent of dimension and diameter [16]. Direct evaluation on such a subset still takes $O(d\varepsilon^{-2})$ time. For fixed $d > 1$, Tai obtains size $O_d(\varepsilon^{-1})$ [17]; the hidden dependence on dimension does not yield the high-dimensional bound sought here. Recent work on query-oblivious softmax-attention coresets likewise uses a simultaneous query guarantee and a different normalized kernel sum [18]. In contrast, the present preprocessing supports each independently fixed query, and a fixed batch of s queries by replacing the amplification factor with $O(\log(2s/\delta))$. A query chosen after inspecting the random map is outside that guarantee.

2 The pointwise KDE theorem

Theorem 2.1 (Pointwise Gaussian KDE without diameter dependence). *Let $d \geq 1$, let $\emptyset \neq X \subset \mathbb{R}^d$ be finite, let $\sigma > 0$, and let*

$$k_\sigma(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{\sigma^2}\right).$$

For every $0 < \varepsilon < 1$ and $0 < \delta < 1$, there is a randomized KDE data structure such that, for every fixed query $y \in \mathbb{R}^d$,

$$\mathbb{P}\left[\left|\frac{1}{|X|} \sum_{x \in X} k_\sigma(x, y) - \tilde{E}(y)\right| < \varepsilon\right] > 1 - \delta.$$

Its query time is

$$O((d + \varepsilon^{-2}) \log(D + 1) \log(2/\delta)),$$

where D is the next power of two at least d . It uses $O((d + \varepsilon^{-2}) \log(2/\delta))$ stored real numbers. In particular the query time is $\tilde{O}(d + \varepsilon^{-2})$ and has no dependence on the effective diameter Δ_σ . Its preprocessing time is

$$O(|X|(d + \varepsilon^{-2}) \log(D + 1) \log(2/\delta)).$$

The theorem is proved in Sections 3 and 4. Padding vectors by zeros allows us to work in $\mathbb{R}^D$, where $D = 2^m$ and $d \leq D < 2d$ for $d \geq 1$. The query-time bound counts ordinary arithmetic operations and evaluations of sine and cosine. As in standard random-feature KDE structures, the data structure stores the empirical feature mean.

3 A fast Gaussian block

Let H be the normalized $D \times D$ Walsh–Hadamard matrix. Index rows and columns by the group $\mathbb{F}_2^m$, so

$$H_{ij} = D^{-1/2} \chi_i(j), \quad \chi_i(j) = (-1)^{i \cdot j}.$$

Let $B = \text{diag}(b_r)$, where the b_r are independent random signs, and let $G = \text{diag}(g_j)$, where the g_j are independent standard Gaussian random variables. All variables in B and G are mutually independent. Define

$$W = \sqrt{D} H G H B.$$

For a fixed displacement $z \in \mathbb{R}^D$, define the one-block estimator

$$Y(z) = \frac{1}{D} \sum_{i \in \mathbb{F}_2^m} \cos((Wz)_i).$$

The goal of this section is to prove

$$\mathbb{E}Y(z) = e^{-\|z\|_2^2/2}, \quad \text{Var}(Y(z)) \leq \frac{5}{D}.$$

Lemma 3.1 (Row marginals). *For every row index i , conditional on B , the i th row w_i of W is distributed as $N(0, I_D)$.*

Proof. Let h_j^T denote the j th row of H . The vectors $h_j^T B$ form an orthonormal basis, because H is orthogonal and B is diagonal with signs. The i th row of W is

$$w_i^T = \sqrt{D} e_i^T H G H B = \sum_{j \in \mathbb{F}_2^m} \chi_i(j) g_j h_j^T B.$$

This is a Gaussian linear combination of an orthonormal basis with independent $N(0, 1)$ coefficients, up to deterministic signs. Hence $w_i \sim N(0, I_D)$ conditional on B . $\square$

Lemma 3.2 (Walsh covariance parameters). *Fix $z \in \mathbb{R}^D$, set $t = \|z\|_2$, and put $a = HBz$. For $R_i = (Wz)_i$, conditional on B the vector $R = (R_i)_i$ is centered Gaussian, $\text{Var}(R_i | B) = t^2$, and for $i \neq k$,*

$$\text{Cov}(R_i, R_k | B) = \rho_{i+k},$$

where

$$\rho_\ell = \sum_{r \in \mathbb{F}_2^m} b_r b_{r+\ell} z_r z_{r+\ell}.$$

For $\ell = 0$ this formula gives $\rho_0 = t^2$.

Proof. Since $a = HBz$,

$$R_i = \sqrt{D} \sum_j H_{ij} g_j a_j.$$

Thus

$$\text{Cov}(R_i, R_k | B) = D \sum_j H_{ij} H_{kj} a_j^2.$$

For $i = k$ this is $\sum_j a_j^2 = \|z\|_2^2$, because HB is orthogonal. For $i \neq k$, write $\ell = i + k$ in $\mathbb{F}_2^m$. Expanding $a_j = D^{-1/2} \sum_r \chi_j(r) b_r z_r$ gives

$$\begin{aligned} D \sum_j H_{ij} H_{kj} a_j^2 &= \sum_j \chi_{i+k}(j) a_j^2 \\ &= D^{-1} \sum_{r,s} b_r b_s z_r z_s \sum_j \chi_j(\ell + r + s). \end{aligned}$$

Walsh orthogonality makes the inner sum equal to D exactly when $s = r + \ell$ and zero otherwise. The displayed formula follows. $\square$

Lemma 3.3 (Cosine covariance identity). *For centered jointly Gaussian random variables U, V with common variance t^2 and covariance ρ ,*

$$\text{Cov}(\cos U, \cos V) = e^{-t^2} (\cosh \rho - 1).$$

Consequently,

$$\text{Var}(Y(z)) = \frac{(1 - e^{-t^2})^2}{2D} + \frac{e^{-t^2}}{D} \mathbb{E}_B \sum_{\ell \neq 0} (\cosh \rho_\ell - 1).$$

Proof. The identity $\cos u \cos v = (\cos(u+v) + \cos(u-v))/2$ and the formula $\mathbb{E} \cos Z = e^{-\text{Var}(Z)/2}$ for centered Gaussian Z give

$$\mathbb{E}[\cos U \cos V] = \frac{1}{2} e^{-(2t^2+2\rho)/2} + \frac{1}{2} e^{-(2t^2-2\rho)/2} = e^{-t^2} \cosh \rho.$$

Also $\mathbb{E} \cos U = \mathbb{E} \cos V = e^{-t^2/2}$, proving the covariance identity.

Condition on B and apply the identity to the Gaussian vector $R = (Wz)_i$. For each fixed ℓ , exactly D ordered pairs (i, k) satisfy $i + k = \ell$. Therefore

$$\text{Var}_G(Y(z) | B) = \frac{e^{-t^2}}{D} \sum_\ell (\cosh \rho_\ell - 1).$$

The conditional mean is $e^{-t^2/2}$ by Lemma 3.1, independent of B , so the law of total variance introduces no additional term. The $\ell = 0$ term is

$$\frac{e^{-t^2}}{D} (\cosh t^2 - 1) = \frac{(1 - e^{-t^2})^2}{2D},$$

which gives the formula. $\square$

Lemma 3.4 (Signed autocorrelation moment). *For fixed $z \in \mathbb{R}^D$ and $t = \|z\|_2$,*

$$|\rho_\ell| \leq t^2 \quad \text{for every } \ell,$$

and

$$\mathbb{E}_B \sum_{\ell \neq 0} \rho_\ell^2 = 2(t^4 - \|z\|_4^4) \leq 2t^4.$$

Proof. The deterministic bound follows from Cauchy–Schwarz:

$$|\rho_\ell| \leq \sum_r |z_r| |z_{r+\ell}| \leq \left(\sum_r z_r^2 \right)^{1/2} \left(\sum_r z_{r+\ell}^2 \right)^{1/2} = t^2.$$

For $\ell \neq 0$, expand

$$\rho_\ell^2 = \sum_{r,s} z_r z_{r+\ell} z_s z_{s+\ell} b_r b_{r+\ell} b_s b_{s+\ell}.$$

The expectation over independent signs is nonzero exactly when the unordered pairs $\{r, r+\ell\}$ and $\{s, s+\ell\}$ coincide, which happens for $s = r$ or $s = r + \ell$. Thus

$$\mathbb{E}_B \rho_\ell^2 = 2 \sum_r z_r^2 z_{r+\ell}^2.$$

Summing over nonzero ℓ gives

$$\mathbb{E}_B \sum_{\ell \neq 0} \rho_\ell^2 = 2 \sum_r z_r^2 \sum_{\ell \neq 0} z_{r+\ell}^2 = 2 \sum_r z_r^2 (t^2 - z_r^2) = 2(t^4 - \|z\|_4^4).$$

□

Lemma 3.5 (Gaussian damping). *For every $a \geq 0$ and every real u with $|u| \leq a$,*

$$e^{-a}(\cosh u - 1) \leq \frac{2u^2}{(1+a)^2}.$$

Proof. The case $a = 0$ is immediate. Assume $a > 0$. The identity

$$\cosh s - 1 = s^2 \int_0^1 (1-r) \cosh(rs) dr$$

shows that $(\cosh s - 1)/s^2$ is nondecreasing for $s > 0$. Hence, for $|u| \leq a$,

$$\cosh u - 1 \leq u^2 \frac{\cosh a - 1}{a^2}.$$

Multiplying by e^{-a} and using

$$e^{-a}(\cosh a - 1) = \frac{(1 - e^{-a})^2}{2}$$

gives

$$e^{-a}(\cosh u - 1) \leq \frac{u^2}{2} \left(\frac{1 - e^{-a}}{a} \right)^2.$$

If $0 < a \leq 1$, then $(1 - e^{-a})/a \leq 1 \leq 2/(1+a)$. If $a \geq 1$, then $(1 - e^{-a})/a \leq 1/a \leq 2/(1+a)$. This proves the displayed inequality. □

Proposition 3.6 (One-block variance). *For every fixed $z \in \mathbb{R}^D$,*

$$\mathbb{E}Y(z) = e^{-\|z\|_2^2/2}, \quad \text{Var}(Y(z)) \leq \frac{5}{D}.$$

Proof. The expectation follows from Lemma 3.1: each row of W is marginally $N(0, I_D)$, so each summand has expectation $e^{-\|z\|_2^2/2}$.

For the variance, use Lemma 3.3. The diagonal term is at most $1/(2D)$. For the off-diagonal term, Lemma 3.4 gives $|\rho_\ell| \leq t^2$, so Lemma 3.5, with $a = t^2$ and $u = \rho_\ell$, gives

$$\begin{aligned} \frac{e^{-t^2}}{D} \mathbb{E}_B \sum_{\ell \neq 0} (\cosh \rho_\ell - 1) &\leq \frac{2}{D(1+t^2)^2} \mathbb{E}_B \sum_{\ell \neq 0} \rho_\ell^2 \\ &\leq \frac{4t^4}{D(1+t^2)^2} \leq \frac{4}{D}. \end{aligned}$$

Thus $\text{Var}(Y(z)) \leq 1/(2D) + 4/D \leq 5/D$. □

4 From pairwise estimation to KDE

For the original bandwidth σ , pad every vector by zeros and set

$$u_x = \sqrt{2}x/\sigma \in \mathbb{R}^D.$$

For one block W , define the feature map

$$\phi_W(x) = \frac{1}{\sqrt{D}} (\cos(Wu_x), \sin(Wu_x)) \in \mathbb{R}^{2D}.$$

Then

$$\phi_W(x)^T \phi_W(y) = \frac{1}{D} \sum_{i=1}^D \cos((W(u_x - u_y))_i) = Y(u_x - u_y).$$

By Proposition 3.6, this is an unbiased estimator of $k_\sigma(x, y)$.

Lemma 4.1 (KDE variance lift). *For one block and a fixed query y , define*

$$\hat{E}_1(y) = \frac{1}{|X|} \sum_{x \in X} Y(u_x - u_y).$$

Then

$$\mathbb{E} \hat{E}_1(y) = \frac{1}{|X|} \sum_{x \in X} k_\sigma(x, y), \quad \text{Var}(\hat{E}_1(y)) \leq \frac{5}{D}.$$

Proof. The expectation follows from

$$e^{-\|u_x - u_y\|_2^2/2} = e^{-\|x - y\|_2^2/\sigma^2} = k_\sigma(x, y).$$

For the variance, write $p_x = 1/|X|$ and

$$Z_x = Y(u_x - u_y) - e^{-\|u_x - u_y\|_2^2/2}.$$

Then

$$\hat{E}_1(y) - \mathbb{E} \hat{E}_1(y) = \sum_{x \in X} p_x Z_x.$$

Minkowski's inequality in L_2 gives

$$\left\| \sum_x p_x Z_x \right\|_{L_2} \leq \sum_x p_x \|Z_x\|_{L_2} \leq \sum_x p_x \sqrt{5/D} = \sqrt{5/D}.$$

Squaring gives the variance bound. No independence among the Z_x is required. □

Proof of Theorem 2.1. Let

$$q = \max \left\{ 1, \left\lceil \frac{20}{\varepsilon^2 D} \right\rceil \right\}.$$

One constant-success copy of the data structure uses q independent blocks $W_1, \dots, W_q$, each with fresh mutually independent sign and Gaussian diagonals, and the concatenated feature map

$$\phi(x) = \frac{1}{\sqrt{q}} (\phi_{W_1}(x), \dots, \phi_{W_q}(x)).$$

The stored vector is the empirical feature mean

$$F_X = \frac{1}{|X|} \sum_{x \in X} \phi(x).$$

On query y , the copy returns $F_X^T \phi(y)$, equivalently the average of the q independent one-block KDE estimates.

By Lemma 4.1 and independence of the blocks,

$$\text{Var}(F_X^T \phi(y)) \leq \frac{5}{qD} \leq \frac{\varepsilon^2}{4}.$$

Chebyshev's inequality gives failure probability at most $1/4$ for additive error ε . Choose an odd integer $L \geq C \log(2/\delta)$, for a sufficiently large absolute constant C , use L mutually independent constant-success copies, and return the median of their estimates. A standard Chernoff bound shows that the median fails with probability at most $\delta/2$. In particular, its success probability is at least $1 - \delta/2 > 1 - \delta$, as required by the strict inequality in the source definition.

For one block, computing $Wu_y = \sqrt{D} HGH B u_y$ uses two fast Walsh–Hadamard transforms and diagonal multiplications, for time $O(D \log D)$. The trigonometric evaluations and inner-product contribution cost $O(D)$. Since

$$qD \leq D + 20\varepsilon^{-2}$$

and $D < 2d$ for $d \geq 1$, one constant-success copy has query time

$$O((d + \varepsilon^{-2}) \log(D + 1)).$$

Median amplification multiplies this by $O(\log(2/\delta))$. The concatenated feature mean and the random diagonal entries require $O((d + \varepsilon^{-2}) \log(2/\delta))$ stored real numbers. Computing and accumulating these features for every point of X takes

$$O(|X|(d + \varepsilon^{-2}) \log(D + 1) \log(2/\delta))$$

preprocessing time. The construction and proof never use a diameter bound on the data or query region, so no Δ_σ dependence appears. $\square$

Corollary 4.2 (Fixed finite query batches). *Let $y_1, \dots, y_s$ be fixed independently of the preprocessing randomness. Running the data structure with per-query failure probability δ/s answers all s queries correctly with probability at least $1 - \delta$, with per-query time multiplied only by the factor $O(\log(2s/\delta))$ in the median step.*

Proof. Apply Theorem 2.1 to each fixed y_j with failure probability δ/s and take a union bound. $\square$

5 Why the statement is pointwise

The proof above uses cancellation in the random signs B for each fixed displacement z before the preprocessing randomness is sampled. It does not imply that a single realized block has small variance uniformly over all possible directions. The following calculation makes the distinction explicit.

Proposition 5.1 (A one-block nonuniformity witness). *Fix a realization of the sign matrix B in one Fastfood block, leaving G random. For any $t > 0$, define $z \in \mathbb{R}^D$ by*

$$z_r = t b_r / \sqrt{D}.$$

Then $\|z\|_2 = t$, every nonzero signed autocorrelation satisfies $\rho_\ell = t^2$, and

$$Y(z) = \cos(tg_0),$$

where g_0 is the Gaussian diagonal entry indexed by the trivial Walsh character.

Proof. The norm identity is immediate. For $\ell \neq 0$,

$$\rho_\ell = \sum_r b_r b_{r+\ell} \left(t b_r / \sqrt{D} \right) \left(t b_{r+\ell} / \sqrt{D} \right) = t^2.$$

Also $Bz = (t/\sqrt{D})\mathbf{1}$, so $HBz = te_0$, where e_0 is the coordinate of the trivial Walsh character. Therefore

$$Wz = \sqrt{D} HG(te_0) = \sqrt{D} tg_0 H e_0 = tg_0 \mathbf{1},$$

and every cosine summand equals $\cos(tg_0)$. In particular,

$$\text{Var}_G(Y(z)) = \frac{1}{2} \left(1 - e^{-t^2} \right)^2,$$

which is bounded away from zero for fixed $t > 0$ independently of D . □

Thus the theorem should not be read as a uniform guarantee over a continuum of queries after the randomness is fixed, nor as an adaptive guarantee for queries chosen after inspecting the random structure or as a function of earlier answers. This matches the pointwise quantifiers in [1, Def. 1.1]. Proposition 5.1 concerns one realized block only. It is not a lower bound for the average of independent blocks or for the median-amplified data structure in Theorem 2.1; it only shows that the signed autocorrelation cancellation cannot be upgraded to a uniform one-block claim.

References

- [1] Tal Wagner. New bounds for kernel sums via fast spherical embeddings. In *Proceedings of the 43rd International Conference on Machine Learning*, PMLR 306:124205–124223, 2026. Published version. The open question is in Section 6; references to Section 7 use the preprint. [arXiv:2605.01263v1 \[cs.DS\]](#), 2 May 2026.
- [2] Arturs Backurs, Zinan Lin, Sepideh Mahabadi, Sandeep Silwal, and Jakub Tarnawski. Efficiently computing similarities to private datasets. In *The Twelfth International Conference on Learning Representations*, 2024. [OpenReview:HMe5CJv9dQ](#).
- [3] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pages 1177–1184, 2007. [NeurIPS proceedings record](#).

- [4] L. Greengard and J. Strain. The fast Gauss transform. *SIAM J. Sci. Stat. Comput.*, 12(1):79–94, 1991. <https://doi.org/10.1137/0912004>.
- [5] M. Charikar and P. Siminelakis. Hashing-based-estimators for kernel density in high dimensions. In *Proc. 58th IEEE Symposium on Foundations of Computer Science (FOCS)*, 1032–1043, 2017. <https://doi.org/10.1109/FOCS.2017.99>.
- [6] M. Charikar, M. Kapralov, N. Nouri, and P. Siminelakis. Kernel density estimation through density constrained near neighbor search. In *Proc. 61st IEEE Symposium on Foundations of Computer Science (FOCS)*, 172–183, 2020. <https://doi.org/10.1109/FOCS46700.2020.00025>.
- [7] N. Ailon and B. Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009. <https://doi.org/10.1137/060673096>.
- [8] Quoc V. Le, Tamás Sarlós, and Alexander J. Smola. Fastfood—computing Hilbert space expansions in loglinear time. In *Proceedings of the 30th International Conference on Machine Learning*, PMLR 28(3):244–252, 2013. PMLR:v28/le13.
- [9] Chang Feng, Qinghua Hu, and Shizhong Liao. Random feature mapping with signed circulant matrix projection. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 3490–3496, 2015. IJCAI proceedings record.
- [10] Krzysztof Choromanski and Vikas Sindhwani. Recycling randomness with structure for sublinear time kernel expansions. In *Proceedings of the 33rd International Conference on Machine Learning*, PMLR 48:2502–2510, 2016. PMLR:v48/choromanski16.
- [11] Quoc Viet Le, Tamás Sarlós, and Alexander J. Smola. Fastfood: approximate kernel expansions in loglinear time. Expanded version, [arXiv:1408.3060v1](https://arxiv.org/abs/1408.3060), 2014.
- [12] Nizar Demni and Hachem Kadri. Orthogonal random features: explicit forms and sharp inequalities. *Transactions on Machine Learning Research*, September 2024. OpenReview:FMtRZ4xzSi.
- [13] Felix Xinnan Yu, Ananda Theertha Suresh, Krzysztof Choromanski, Daniel Holtmann-Rice, and Sanjiv Kumar. Orthogonal random features. In *Advances in Neural Information Processing Systems*, volume 29, 2016. NeurIPS proceedings record.
- [14] Krzysztof Choromanski, Mark Rowland, and Adrian Weller. The unreasonable effectiveness of structured random orthogonal embeddings. In *Advances in Neural Information Processing Systems*, volume 30, 2017. NeurIPS proceedings record.
- [15] Krzysztof Choromanski, Mark Rowland, Tamás Sarlós, Vikas Sindhwani, Richard E. Turner, and Adrian Weller. The geometry of random features. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, PMLR 84:1–9, 2018. PMLR:v84/choromanski18a.
- [16] Jeff M. Phillips and Wai Ming Tai. Improved coresets for kernel density estimates. In *Proceedings of the 29th Annual ACM–SIAM Symposium on Discrete Algorithms*, pages 2718–2727, 2018. doi:10.1137/1.9781611975031.173.
- [17] Wai Ming Tai. Optimal coreset for Gaussian kernel density estimation. In *38th International Symposium on Computational Geometry*, LIPIcs 224, pages 63:1–63:15, 2022. doi:10.4230/LIPIcs.SoCG.2022.63.
- [18] Ofek I. Cohen. Query-oblivious coresets for softmax attention: improved bounds and efficient constructions. [arXiv:2609.06327v2](https://arxiv.org/abs/2609.06327), 13 September 2026.