
Bernstein–Schur Kernels: Random Features by Sketched Modulation and Radial Randomization

Taha Bouhsine
Azetta AI

taha@azetta.ai

Abstract

Bernstein–Schur kernels are products of a finite-feature kernel and a completely monotone shift-invariant kernel: nonstationary kernels falling between the shift-invariant and dot-product templates random features exploit, so neither Bochner sampling nor polynomial sketching applies to the full kernel directly. We give one random-feature construction for the whole class that *randomizes both factors*: it sketches the finite modulation and samples the radial factor’s one-dimensional Bernstein–Widder scale before applying Gaussian random Fourier features, giving feature dimension Dm , free of the $O(d^2)$ size of the exact modulation feature. With the modulation kept exact (the $m \rightarrow \infty$ limit), we prove unbiasedness, an exact variance, and a matrix-Bernstein operator-norm bound controlled by the top kernel and modulation eigenvalues and an intrinsic dimension rather than the crude $N \max_{ij}$ route. Whitening this argument at the ridge makes the effective dimension $d_{\text{eff}}(\lambda)$ the *exact* intrinsic dimension of the matrix variance, so $O((1 + \|P\|_{\text{op}}/\lambda) \log(d_{\text{eff}}/\delta))$ radial draws preserve the kernel-ridge solution; tilting the draw by a closed-form whitened leverage improves this to the effective-dimension count $O((1 + d_{\text{eff}}) \log(d_{\text{eff}}/\delta))$. Conditioning on the sketch carries every guarantee to the deployed doubly-randomized estimator up to one additive sketch term, and all hold for the whole class with the modulation Gram in place of the polynomial one. The flagship instance is the biased \mathbf{E} -kernel $k_{\mathbf{E},b}(w, x) = (w^\top x + b)^2 / (\|w - x\|^2 + \varepsilon)$, whose family span contains the inverse-multiquadric kernel by finite differences in b . Experiments validate the construction off-sphere, where the kernel is genuinely non-dot-product and Nyström degrades with d at matched landmark count, and show on a controlled test that the \mathbf{E} -kernel is preferred on the coupled alignment–proximity target and not on single-factor controls.

1 Introduction

Kernel methods (support vector machines (Cortes & Vapnik, 1995), kernel ridge regression, Gaussian processes (Rasmussen & Williams, 2006), and the attention layers of modern transformers, which compute a kernel smoother over tokens (Tsai et al., 2019)) rest on the $N \times N$ Gram matrix $K_{ij} = k(x_i, x_j)$, whose $O(N^2)$ storage and $O(N^3)$ factorization become infeasible once N exceeds 10^5 . The kernel may be fixed or learned implicitly (Li et al., 2019), but in every case scaling these methods means never forming K explicitly, and the dominant tool is the random feature map: a low-dimensional z with $\mathbb{E}[z(x)^\top z(w)] = k(x, w)$, so that $K \approx ZZ^\top$ is handled in $O(NM)$ time and memory. For attention in particular, this is the random-feature linear-attention route (Katharopoulos et al., 2020; Choromanski et al., 2021).

Which kernels admit such a map is dictated by their structure. *Random Fourier features* (RFF) (Rahimi & Recht, 2007) cover shift-invariant kernels $k(x, w) = \phi(x - w)$ through Bochner’s theorem, sampling the nonnegative spectral measure whose Fourier transform is ϕ ; quasi-Monte Carlo (Avron et al., 2016) and orthogonal frequencies (Yu et al., 2016) reduce their variance, and the construction underpins large-scale kernel ridge regression (Avron et al., 2017). *Polynomial sketches* and dot-product random features (TensorSketch, Fastfood, and the dot-product feature maps of Pham & Pagh, 2013; Le et al., 2013; Kar & Karnick, 2012, with oblivious high-degree sketches by Ahle et al., 2020 and complex-valued, variance-optimized improvements by Wacker et al., 2024) do the same for dot-product kernels $k(x, w) = \kappa(x^\top w)$. *Nyström methods*

(Williams & Seeger, 2000; Musco & Musco, 2017) take a different route, building a low-rank approximation from $m \ll N$ landmark columns at the mercy of the kernel’s spectral decay. Each family is bound to a structure: RFF to shift-invariance, sketches to dot-product form, Nyström to fast eigenvalue decay. A unifying thread is that any kernel written as a *mixture* can be linearized by sampling the mixture, which extends RFF beyond Gaussian spectra to broad stationary families (Wilson & Adams, 2013); our radial factor is handled in exactly this way. What is specific here is that the mixture is applied to only the radial part of a kernel that is *not* stationary overall, and is composed with an exact finite feature for the polynomial part, so the scheme sits at the intersection of mixture-based RFF and exact feature maps rather than inside either.

The kernel we wish to scale fits none of these molds. The biased \mathbf{E} -kernel

$$k_{\mathbf{E},b}(w, x) = \frac{(w^\top x + b)^2}{\|w - x\|^2 + \varepsilon}, \quad b \geq 0, \varepsilon > 0, \quad (1)$$

couples alignment and proximity: a squared inner product in the numerator, an inverse squared-distance in the denominator. The finite-difference identity in Proposition A.11 shows that the family $\text{span}\{k_{\mathbf{E},b} : b \geq 0\}$ contains the inverse-multiquadric (IMQ) kernel. We do not rely on any single-kernel universality statement in this paper; fixed- $b > 0$ universality is discussed separately by Bouhsine (2026). In the construction and experiments we treat a fixed $b > 0$ as a practical inductive-bias parameter, and the unbiased $k_{\mathbf{E}}$ ($b = 0$), whose RKHS functions all vanish at the origin, is the special case. But $k_{\mathbf{E},b}$ is *not* shift-invariant and *not* a dot-product kernel: the radial denominator $\|x - w\|^2$ alone is shift-invariant, yet multiplying it by the alignment numerator $(x^\top w + b)^2$ destroys both forms at once, so neither Bochner sampling nor polynomial sketching applies to the full kernel, and exact methods revert to the $O(N^2)$ Gram matrix. (On the unit sphere $\|x - w\|^2 = 2 - 2x^\top w$, so $k_{\mathbf{E},b}$ coincides with a rational dot-product kernel and a dimension-efficient dot-product route becomes available; deriving that route is left to future work, and we target the general \mathbb{R}^d construction here, where no dot-product reduction of the *full* kernel exists. One weaker factorization does survive off-sphere: at each fixed scale the Gaussian factors through positive per-point envelopes, $e^{-t\|x-w\|^2} = e^{-2t\|x\|^2} e^{-2t\|w\|^2} \mathbb{E}[\dots]$, the FAVOR⁺ route of Proposition D.1; but composed with the natural mixing law it has *infinite variance* on every pair with $x^\top w \geq \varepsilon/8$ (Proposition D.2), so the trigonometric construction below is not one choice among equals.) The key empirical test is therefore off-sphere (Figure 1): on a bounded ball with varying norms (Section 5.1) the kernel is genuinely non-dot-product, and RAY (Random Approximation of the \mathbf{E} -kernel) retains its $O(1/\sqrt{D})$ behavior while landmark methods degrade with d .

The obstacle dissolves once $k_{\mathbf{E},b}$ is read as a Schur product (Eq. 2) of two individually tractable pieces: the biased degree-2 polynomial kernel $(w^\top x + b)^2$, which has an exact finite feature map, and the IMQ-type radial kernel $(\|w - x\|^2 + \varepsilon)^{-1}$, which, being completely monotone in $\|x - w\|^2$, is a Bernstein–Widder mixture of Gaussians over a single scale parameter (the classical Schoenberg radial kernels (Schoenberg, 1938), for which broad scale-mixture samplers exist beyond this IMQ case (Langrené et al., 2024)). We turn this into a feature scheme that randomizes *both* factors: sample the radial factor’s one-dimensional scale, apply random Fourier features to the resulting Gaussian, and sketch the polynomial factor. Keeping the polynomial exact is the analyzable limit ($m \rightarrow \infty$), where the variance and concentration are sharpest; but that feature carries the $O(d^2)$ polynomial size, whereas sketching makes the dimension Dm , free of d^2 , at the cost of one controllable sketch term (Theorem 4.4). Beyond the radial random Fourier features, the only added randomness is the one-dimensional Bernstein scale, so for a fixed dataset the radial draw count carries no explicit d -dependence (as for standard RFF), the genuine contrast with uniform Nyström, which we confirm degrades with d .

The point is not that products of kernels can be tensored, a standard closure property, but that factoring this particular nonstationary, non-dot-product kernel turns each piece into a textbook object (an exact finite polynomial feature and a Gaussian mixture), so a kernel fitting neither random-feature template is linearized by combining the standard tools for each. This places RAY among random-feature methods that reach beyond Bochner sampling: for dot-product (Kar & Karnick, 2012; Pham & Pagh, 2013), compositional (Daniely et al., 2017), and non-stationary spectral or harmonizable (Remes et al., 2017; Ton et al., 2018; Shen et al., 2019) kernels, from which it differs by factoring the kernel into a finite modulation and a completely monotone radial factor and randomizing each by its native tool, a polynomial sketch and a Bernstein–Widder radial sampler (Section 6).

The contribution is the analysis of the resulting estimator and the class it defines. For exact modulation we prove unbiasedness, an exact variance for the flat estimator, an expected matrix-Bernstein operator-norm bound (with a matching tail) controlled by the top eigenvalues of the kernel and modulation Gram matrices together with an intrinsic dimension, and relative-spectral kernel-ridge stability; conditioning on the sketch carries the operator-norm guarantee to the deployed doubly-randomized estimator. Unbiasedness, the variance, and the uniform entrywise bound lift verbatim to the *Bernstein–Schur* class (a finite-feature kernel times a completely monotone radial kernel, of which $k_{\mathbf{E},b}$ is the flagship) by a single theorem (Theorem 3.6); the matrix-Bernstein and kernel-ridge statements, proved for $k_{\mathbf{E},b}$, extend to every member with the modulation Gram in place of the polynomial one (Theorem 4.13). The construction is self-contained: positive-definiteness from the Schur factorization (Eq. 2), the IMQ span from the finite-difference identity (Proposition A.11), and the radial mixture from Bernstein–Widder (Section 3.2); we rely on Bouhsine (2026) only for the historical infinite-feature interpretation and fixed- b universality not needed here.

Contributions.

- (i) **A random-feature construction for a kernel class, not one kernel.** We identify *Bernstein–Schur kernels*, a finite-feature kernel times a completely monotone shift-invariant kernel, and give one estimator for the whole class: keep the finite feature exact, sample the radial Bernstein–Widder scale, and apply random Fourier features (Theorem 3.6, unbiased with variance and uniform bounds). The biased \mathbf{E} -kernel is the flagship instance (Section 3, Proposition 3.1 shows it is genuinely neither stationary nor dot-product); the deployed estimator randomizes both factors, with exact modulation as the analyzable limit (Section 3.5).
- (ii) **Sharp variance and optimal allocation** (Section 4): the *exact* variance of the recommended flat estimator (Theorem 4.1), a two-level identity proving one frequency per scale is variance-optimal within the hierarchical estimator at a fixed radial-feature budget (Proposition A.3), and a normalized variant whose variance is bounded free of the data radius and bias (Proposition A.8).
- (iii) **Gram concentration and high-probability KRR guarantees:** a uniform Gram bound (Theorem A.4) and an expected matrix-Bernstein operator-norm bound (Theorem 4.2) with a matching high-probability tail (Corollary 4.3), controlled by the top eigenvalues of the kernel and modulation Gram matrices and an intrinsic dimension rather than the crude $N \max_{ij}$ entrywise route (while still allowing worst-case N -scaling through those spectra). Whitening the same argument at the ridge $A = K + \lambda I$ makes the ridge effective dimension $d_{\text{eff}}(\lambda) = \text{tr}(K(K + \lambda I)^{-1})$ the *exact* intrinsic dimension of the matrix variance (Theorem 4.8): $D = O((1 + \|P\|_{\text{op}}/\lambda) \log(d_{\text{eff}}/\delta))$ radial draws give the relative-spectral condition of Theorem 4.7 with high probability, hence coefficient stability and objective-value preservation to constants (Corollary 4.9), carried to the deployed sketched estimator (Corollary 4.12). Tilting the radial draw by a closed-form *whitened leverage* (a quadratic form whose mean over the base law is exactly $d_{\text{eff}}(\lambda)$) replaces the $\|P\|_{\text{op}}/\lambda$ factor by $d_{\text{eff}}(\lambda)$ itself, the leverage-sampled count of the random-feature KRR literature, with the same proof (Theorem 4.10). All matrix-level guarantees hold for the whole Bernstein–Schur class with the modulation Gram in place of P (Theorem 4.13); a full excess-risk theorem additionally needs control of the approximate kernel’s effective dimension, which we leave open (Remark 4.11).
- (iv) **RAY, the doubly-randomized estimator, and the signal-modulation decomposition.** The exact polynomial feature costs $O(d^2)$ per draw, so RAY sketches it: the feature dimension becomes Dm , free of d^2 . Conditioning on the sketch, this doubly-randomized estimator carries the *same* intrinsic-dimension operator-norm guarantee as the exact-modulation limit plus a single additive sketch term, tunable by m independently of D (Theorem 4.4); exact modulation is the $m \rightarrow \infty$ limit. A modulation–radial error decomposition (Proposition 4.14) explains why the product still helps: radial approximation noise is modulated by the true finite feature, modulation error is localized by radial proximity, and only an interaction term mixes them, so the alignment \times proximity product suppresses similarity false positives even when the modulation is sketched (Appendix D.1).
- (v) **Empirical validation** (Section 5): an off-sphere bounded-ball test confirming the general- \mathbb{R}^d niche where no direct dot-product reduction is available and Nyström degrades with d (Section 5.1), a Ten-

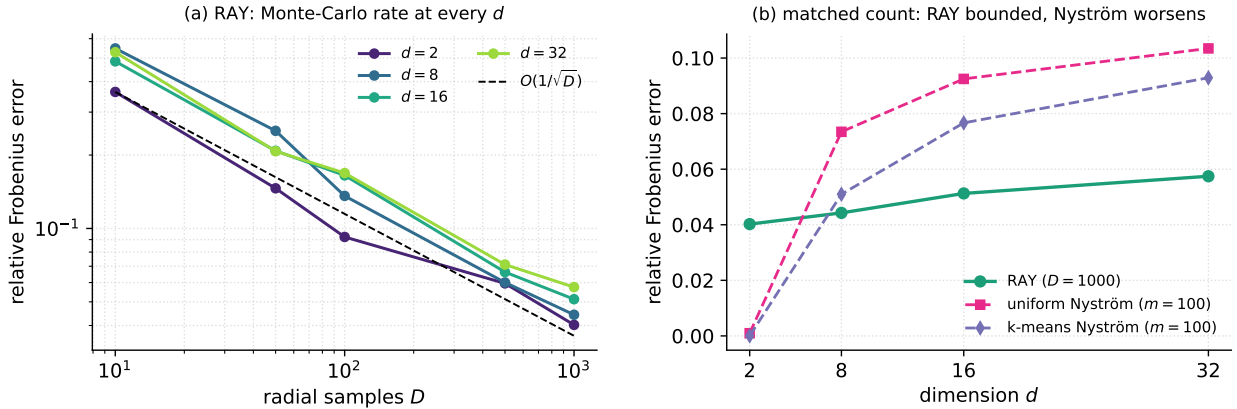


Figure 1: The key regime: an *off-sphere* bounded ball (varying norms), where $k_{\mathbf{E},b}$ is genuinely non-dot-product and no direct dot-product reduction is available. (a) RAY’s relative Frobenius Gram error follows the $O(1/\sqrt{D})$ Monte-Carlo rate at every dimension. (b) At $D = 1000$ RAY stays bounded as d grows, while uniform and k-means Nyström (fixed $m = 100$ landmarks) degrade (matched in radial/landmark count; the cost-matched comparison is Appendix D.1).

sorSketch deployment study showing the radial-plus-sketch error split and removal of the $O(d^2)$ modulation floor (Section 5.2), a controlled coupled-target preference test where the \mathbf{E} -kernel is the best bias exactly on the alignment–proximity target and not on single-factor controls (Section 5.3), and a large-scale streaming primal run where the Gram is impossible (Section 5.4). Cost-matched caveats, variance/bias checks, downstream KRR, attention, and sphere-normalized sanity checks are deferred to Appendix D.

Section 2 fixes notation and the random-feature background; Section 3 constructs the estimator step by step; Section 4 establishes its guarantees and variance reduction; Section 5 reports the experiments; Sections 6–7 discuss and conclude. Proofs and further validations are in the appendix.

2 Preliminaries

2.1 Notation

Vectors are columns in \mathbb{R}^d ; $x^\top w$ is the inner product, $\|x\|$ the Euclidean norm, and $X = \{x_1, \dots, x_N\} \subset \mathbb{R}^d$ a compact dataset with $\|x_i\| \leq R$. For a kernel k , $K = [k(x_i, x_j)] \in \mathbb{R}^{N \times N}$ is its Gram matrix; \circ is the entrywise (Hadamard/Schur) product, so $(A \circ B)_{ij} = A_{ij}B_{ij}$, and $A \preceq B$ the Loewner order. We write $x \otimes x$ for the Kronecker square and $\text{vec}(\cdot)$ for its vectorization, so $\text{vec}(x \otimes x)^\top \text{vec}(w \otimes w) = (x^\top w)^2$. The target is the biased \mathbf{E} -kernel $k_{\mathbf{E},b}$ of (1), with bias $b \geq 0$ and bandwidth $\varepsilon > 0$; it factors (Section 3.1) into the degree-2 modulation $p_b(x, w) = (x^\top w + b)^2$ with feature $p_b(x)$ of dimension $d_b = d(d+1)/2 + d + 1$, and the radial factor $h_\varepsilon(x, w) = (\|x - w\|^2 + \varepsilon)^{-1}$. A random feature map $z : \mathbb{R}^d \rightarrow \mathbb{R}^M$ satisfies $\mathbb{E}[z(x)^\top z(w)] = k(x, w)$; the deployed estimator uses D radial draws $t_j \sim \text{Exp}(\varepsilon)$ with frequencies $\omega_j \sim \mathcal{N}(0, 2t_j I_d)$ and a degree-2 sketch TS_m of width m , giving feature dimension $M = Dm$. We write $\|\cdot\|_{\text{op}}$ for the operator norm and $\text{Exp}(\varepsilon)$ for the exponential law of rate ε .

2.2 Random features and positive-definite kernels

A continuous symmetric k is positive definite if every Gram K is PSD; by Mercer’s theorem it then has a nonnegative spectral decomposition on compact X . Two closure properties drive the construction. *Schur product*: if k_1, k_2 are PD then so is $k_1 k_2$, with $K_1 \circ K_2 \succeq 0$ (Schur, 1911). *Mixtures*: a nonnegative mixture $\int k_t d\mu(t)$ of PD kernels is PD, and sampling $t \sim \mu$ yields an unbiased random feature — the principle behind random Fourier features for shift-invariant kernels (Rahimi & Recht, 2007) and their scale-mixture extensions

(Wilson & Adams, 2013). A function is *completely monotone* in $\|x - w\|^2$ if it is a Laplace transform of a nonnegative measure; by Bernstein–Widder it is then a Gaussian scale mixture (Section 3.2). The quality of a random-feature Gram $\widehat{K} = ZZ^\top$ is measured here by the relative Frobenius error $\|\widehat{K} - K\|_F/\|K\|_F$ and, for downstream stability, by the operator-norm error $\|\widehat{K} - K\|_{\text{op}}$, whose effective scale is set not by N but by an intrinsic dimension of the kernel spectrum (Section 4).

3 Constructing the Random \mathbb{E} -Feature

We seek a feature map $z : \mathbb{R}^d \rightarrow \mathbb{R}^{M_b}$ with $\mathbb{E}[z(x)^\top z(w)] = k_{\mathbb{E},b}(x, w)$, so that the Gram matrix of $X = \{x_1, \dots, x_N\}$ is approximated by ZZ^\top in $O(NM_b)$ rather than $O(N^2)$. We build z in five steps: (i) factor $k_{\mathbb{E},b}$ into a polynomial and a radial kernel (Section 3.1); (ii) write the radial kernel as a nonnegative mixture of Gaussians over a one-dimensional scale (Section 3.2); (iii) discretize the mixture by sampling that scale (Section 3.3); (iv) approximate the Gaussian factor by random Fourier features and tensor with the modulation feature (Section 3.4); and (v) sketch the modulation, so that *both* factors are randomized and the feature dimension is free of the $O(d^2)$ polynomial size (Section 3.5). Steps (i)–(iv) with an exact modulation give the analyzable limit; step (v) is the deployed estimator.

3.1 Step 1: Schur factorization

The kernel splits as a Schur product of a polynomial and a radial kernel,

$$k_{\mathbb{E},b}(w, x) = \underbrace{(w^\top x + b)^2}_{p_b(w, x)} \cdot \underbrace{\frac{1}{\|w - x\|^2 + \varepsilon}}_{h_\varepsilon(w, x)}, \quad (2)$$

both factors positive definite (p_b from its explicit feature map $p_b(x)^\top p_b(w)$ (Proposition 3.2), and h_ε as a nonnegative Bernstein–Widder mixture of Gaussian kernels (Section 3.2)), so $k_{\mathbb{E},b}$ is PSD by the Schur product theorem (Schur, 1911). Being continuous and positive definite, $k_{\mathbb{E},b}$ is a Mercer kernel in the usual integral-operator sense on every compact $X \subset \mathbb{R}^d$, the setting throughout (Section 4 fixes $\|x\| \leq R$); we make no claim on noncompact domains. (This is self-contained; Bouhsine (2026) is cited only for historical context and fixed- b universality results not used here.) The polynomial factor is the tractable one: it has an exact, finite feature map. That the full kernel nonetheless sits outside both random-feature templates is not a matter of degree but a structural fact.

The radial factor is itself recovered from the family by an elementary *forward* second difference in b : $(2h^2)^{-1}[k_{\mathbb{E},b+2h} - 2k_{\mathbb{E},b+h} + k_{\mathbb{E},b}] = h_\varepsilon$ exactly, for any $b \geq 0$, $h > 0$, with all three biases inside the domain $b \geq 0$ (Proposition A.11, Appendix A). This is the only “spans IMQ” statement we use, and it keeps the claim self-contained.

Proposition 3.1 (Off-sphere the kernel is neither stationary nor dot-product). *For $b \geq 0$, $\varepsilon > 0$, the kernel $k_{\mathbb{E},b}$ is not shift-invariant on any domain containing two points of different norm, and not a dot-product kernel on any domain containing two pairs of equal inner product $s = x^\top w \neq -b$ but different distance (at $s = -b$ the numerator vanishes for both pairs, so this value is excluded). (On the unit sphere, where distance and inner product are in bijection, it does reduce to a dot-product kernel.)*

The witnesses are immediate: $k_{\mathbb{E},b}(x, x) = (\|x\|^2 + b)^2/\varepsilon$ varies with $\|x\|$, ruling out shift-invariance; and $x=w=e_1$ versus $x'=\sqrt{2}e_1$, $w'=e_1/\sqrt{2}$ have $x^\top w = x'^\top w' = 1$ but $\|x - w\|^2 = 0 \neq \frac{1}{2} = \|x' - w'\|^2$, so $k_{\mathbb{E},b}$ is not a function of $x^\top w$ alone. This is why neither Bochner sampling nor dot-product sketching applies to the full kernel, and why the off-sphere experiment (Section 5.1) is the key test.

Proposition 3.2 (Biased polynomial feature). *For $b \geq 0$ define*

$$p_b(x) = (\text{vec}(x \otimes x), \sqrt{2b}x, b)^\top \in \mathbb{R}^{d^2+d+1}. \quad (3)$$

Then $p_b(x)^\top p_b(w) = (x^\top w)^2 + 2b(x^\top w) + b^2 = (x^\top w + b)^2$, and $\|p_b(x)\|^2 = (\|x\|^2 + b)^2$.

Proof. $p_b(x)^\top p_b(w) = \text{vec}(x \otimes x)^\top \text{vec}(w \otimes w) + 2b x^\top w + b \cdot b = (x^\top w)^2 + 2b(x^\top w) + b^2$, where the constant coordinate b contributes $b \cdot b = b^2$. Setting $w = x$ gives $\|p_b(x)\|^2 = \|x\|^4 + 2b\|x\|^2 + b^2 = (\|x\|^2 + b)^2$. \square

This finite feature is the object the deployed estimator will *sketch*. We develop the construction in two passes. First, holding the modulation exact, the estimator’s randomness is entirely radial, so the unbiasedness and variance identities are transparent (Section 4); exact modulation is the $m \rightarrow \infty$ analyzable limit. The deployed estimator then sketches this feature (Step 5, Section 3.5), making the feature dimension free of the $O(d^2)$ polynomial size at the price of a single, controllable modulation-sketch term on top of the radial one (Theorem 4.4). Richer modulations (attention-style features or learned anchors) slot into the same two-error decomposition. The feature is real exactly on the kernel’s domain $b \geq 0$ (the entry $\sqrt{2b}$ requires it), consistent with the definition (1). By symmetry the dimension reduces from d^2 to $d(d+1)/2$ using the upper triangle of $x \otimes x$ with $\sqrt{2}$ scaling on off-diagonals; for large d , TensorSketch (Pham & Pagh, 2013) reduces it further to a controllable D_{poly} (Appendix D). Setting $b = 0$ recovers the unbiased feature $p_0(x) = \text{vec}(x \otimes x)$. From here on, unless stated otherwise, p_b denotes the symmetrized degree-2 feature of dimension $d_b = d(d+1)/2 + d + 1$, exactly equivalent in inner product to the redundant $d^2 + d + 1$ tensor feature of (3).

3.2 Step 2: Bernstein–Widder representation of the radial factor

The radial factor h_ε is completely monotone in $\|x - w\|^2$, so by the Bernstein–Widder theorem (Widder, 1941) it is a Laplace mixture of Gaussian kernels,

$$h_\varepsilon(x, w) = \frac{1}{\|x - w\|^2 + \varepsilon} = \int_0^\infty e^{-t(\|x-w\|^2 + \varepsilon)} dt = \int_0^\infty e^{-t\varepsilon} g_t(x, w) dt, \quad (4)$$

where $g_t(x, w) = e^{-t\|x-w\|^2}$ is the Gaussian kernel at scale t . Multiplying by the polynomial factor and tensoring its feature gives the exact feature map of $k_{\mathbf{E}, b}$ (the infinite-feature interpretation, derived here and discussed historically by Bouhsine, 2026): with $\Phi_b(x)(t) = e^{-t\varepsilon/2} \varphi_t(x) \otimes p_b(x)$ (φ_t the canonical Gaussian feature map),

$$\langle \Phi_b(x), \Phi_b(w) \rangle = \int_0^\infty e^{-t\varepsilon} g_t(x, w) (x^\top w + b)^2 dt = k_{\mathbf{E}, b}(x, w). \quad (5)$$

The mixing measure $\varepsilon e^{-\varepsilon t} dt$ is, up to normalization, the density of $T \sim \text{Exp}(\varepsilon)$. Since $\mathbb{E}_{T \sim \text{Exp}(\varepsilon)}[e^{-T\|x-w\|^2}] = \varepsilon / (\|x - w\|^2 + \varepsilon)$, the kernel is an expectation over that single scale,

$$k_{\mathbf{E}, b}(x, w) = \frac{1}{\varepsilon} \mathbb{E}_{T \sim \text{Exp}(\varepsilon)}[g_T(x, w) (x^\top w + b)^2], \quad (6)$$

and it is this expectation, rather than the integral (5), that the next step discretizes.

3.3 Step 3: Sampling the radial scale

Equation (6) replaces the integral over t by an expectation over a *one-dimensional* random scale. Drawing $t_1, \dots, t_D \stackrel{\text{iid}}{\sim} \text{Exp}(\varepsilon)$ and averaging gives the Monte-Carlo estimate

$$k_{\mathbf{E}, b}(x, w) \approx \frac{(x^\top w + b)^2}{\varepsilon} \cdot \frac{1}{D} \sum_{j=1}^D g_{t_j}(x, w). \quad (7)$$

At this point only the radial scale is sampled and the modulation is still exact, the analyzable limit, whose single (radial) error source makes the unbiasedness and variance identities of Section 4 clean. Step 5 then sketches the modulation, randomizing *both* factors and removing the $O(d^2)$ floor at the cost of one additional, controllable error source (Theorem 4.4). This shapes the scheme’s dimension behavior: the outer Bernstein scale t_j is one-dimensional, and while the inner Gaussian frequency $\omega \in \mathbb{R}^d$ of the radial RFF is still d -dimensional, the fixed-dataset Hoeffding/union-bound count for the radial estimate carries no explicit d -dependence (Corollary A.5), as it does for standard RFF; the genuine d -contrast is in applicability and in the polynomial feature dimension, discussed in Section 4. It remains to make g_{t_j} computable by an inner product.

3.4 Step 4: Random Fourier features and the estimator

Each Gaussian $g_t(x, w) = e^{-t\|x-w\|^2}$ is shift-invariant, so Bochner’s theorem applies to it (even though it does not apply to $k_{\mathbf{E},b}$): drawing $\omega \sim \mathcal{N}(0, 2tI_d)$ and $\beta \sim \text{Unif}([0, 2\pi])$, $\mathbb{E}_{\omega, \beta}[2 \cos(\omega^\top x + \beta) \cos(\omega^\top w + \beta)] = g_t(x, w)$ (Rahimi & Recht, 2007). Tensoring D' such features with the exact polynomial feature yields the estimator; we write $d_b := d(d+1)/2 + d + 1$ for the symmetric polynomial feature (the default; the redundant d^2 tensor is reduced losslessly) and $M_b = DD' d_b$.

Definition 3.3 (Exact-modulation \mathbf{E} -feature). *Given $D, D' \in \mathbb{N}$, $\varepsilon > 0$, $b \geq 0$, the exact-modulation \mathbf{E} -feature map $z : \mathbb{R}^d \rightarrow \mathbb{R}^{M_b}$ (the analyzable limit, which the deployed RAY map of Step 5 randomizes further) is:*

1. Draw $t_1, \dots, t_D \stackrel{\text{iid}}{\sim} \text{Exp}(\varepsilon)$.
2. For each j , draw $\omega_{j,1}, \dots, \omega_{j,D'} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 2t_j I_d)$ and $\beta_{j,\ell} \stackrel{\text{iid}}{\sim} \text{Unif}([0, 2\pi])$ (the factor 2 is the standard RFF convention so that $\mathbb{E}_{\omega, \beta}[2 \cos(\omega^\top x + \beta) \cos(\omega^\top w + \beta)] = e^{-t_j\|x-w\|^2}$).
3. Gaussian RFF at scale t_j : $\psi_{t_j}(x) = \sqrt{2/D'} (\cos(\omega_{j,\ell}^\top x + \beta_{j,\ell}))_{\ell=1}^{D'} \in \mathbb{R}^{D'}$.
4. Exact biased polynomial feature: $p(x) = \varepsilon^{-1/2} p_b(x) \in \mathbb{R}^{d_b}$.
5. Block: $z_j(x) = D^{-1/2} \psi_{t_j}(x) \otimes p(x)$. Concatenate: $z(x) = (z_1(x)^\top, \dots, z_D(x)^\top)^\top$.

The two levels D and D' are less fundamental than they look, and naming what the inner draw actually samples makes this clear. The hierarchical draw $t \sim \text{Exp}(\varepsilon)$, $\omega \sim \mathcal{N}(0, 2tI_d)$ produces ω with marginal density $\int_0^\infty \mathcal{N}(\omega; 0, 2tI_d) \varepsilon e^{-\varepsilon t} dt$, which by the same Gaussian scale mixture (4) is the normalized Bochner spectral distribution of the rescaled radial factor $\varepsilon h_\varepsilon$; the missing total mass $1/\varepsilon$ is carried by the feature scaling. Step 4 is therefore standard random Fourier features for h_ε tensored with the exact polynomial feature. The IMQ spectral density can be written using modified Bessel functions; the radial-scale mixture is simply a tractable sampler for that distribution, not a device that competes with it. Seen this way the inner level is redundant: at equal cost DD' the D' frequencies of a block share one scale t_j and are correlated, whereas D' independent spectral draws are not, so variance is minimized at $D' = 1$ except in the degenerate zero-outer-variance case (Appendix D.3 confirms this directly). We keep D' general for the analysis but recommend $D' = 1$, drawing D independent (t_j, ω_j) pairs, which collapses the estimator to

$$z(x) = D^{-1/2} \left(\sqrt{2} \cos(\omega_j^\top x + \beta_j) p(x) \right)_{j=1}^D, \quad \omega_j \sim \mathcal{N}(0, 2t_j I_d), \quad t_j \sim \text{Exp}(\varepsilon). \quad (8)$$

The contribution of this section is thus the Schur factorization that isolates an exactly-representable, sketchable polynomial modulation (Section 3.1) paired with this sampler for the IMQ factor, not an improvement over IMQ random features; the dimension comparison this invites is taken up in Section 4.2.

Theorem 3.4 (Unbiasedness). $\mathbb{E}[z(x)^\top z(w)] = k_{\mathbf{E},b}(x, w)$.

Proof. $z(x)^\top z(w) = \frac{1}{D} \sum_{j=1}^D (\psi_{t_j}(x)^\top \psi_{t_j}(w)) (p(x)^\top p(w))$. The polynomial inner product is exact: $p(x)^\top p(w) = (x^\top w + b)^2 / \varepsilon$. By the standard RFF identity (Rahimi & Recht, 2007), $\mathbb{E}_\omega[\psi_{t_j}(x)^\top \psi_{t_j}(w) \mid t_j] = e^{-t_j\|x-w\|^2}$. Taking expectation over $t_j \sim \text{Exp}(\varepsilon)$ and using (6),

$$\mathbb{E}[z(x)^\top z(w)] = \mathbb{E}_{T \sim \text{Exp}(\varepsilon)}[e^{-T\|x-w\|^2}] \cdot \frac{(x^\top w + b)^2}{\varepsilon} = \frac{\varepsilon}{\|x-w\|^2 + \varepsilon} \cdot \frac{(x^\top w + b)^2}{\varepsilon} = k_{\mathbf{E},b}(x, w). \quad \square$$

The construction used nothing specific to $k_{\mathbf{E},b}$ beyond its structure as a finite-feature kernel times a completely monotone shift-invariant kernel. The estimator, and its guarantees, hold for the whole class.

Remark 3.5 (Differentiability and learnable (ε, b)). *The estimator is differentiable in its own kernel parameters with the base randomness held fixed. Reparameterize the radial draw as $t_j = -\varepsilon^{-1} \log u_j$, $u_j \sim \text{Unif}[0, 1]$, and $\omega_j = \sqrt{2t_j} g_j$, $g_j \sim \mathcal{N}(0, I_d)$, so ε enters $z(x)$ smoothly through t_j and ω_j , while b enters through the*

exact polynomial feature p_b . At every fixed (ε, b) the map is unbiased (Theorem 3.4), so a gradient of any downstream objective in (ε, b) flows through an unbiased kernel estimate and per-head learnable (ε, b) in \mathbf{E} -attention is a reparameterization away, with ε free to adapt to the local attention sharpness (the hard regime of Appendix D.1).

Theorem 3.6 (Bernstein–Schur random features). *Let $k(x, w) = p(x, w) f(\|x - w\|^2)$ with $p(x, w) = u(x)^\top u(w)$ for a finite feature $u : \mathbb{R}^d \rightarrow \mathbb{R}^{d_p}$, and f completely monotone with Bernstein mixture $f(r) = \int_0^\infty e^{-tr} d\nu(t)$ of finite mass $m_f := \nu(\mathbb{R}_{\geq 0}) = f(0) < \infty$ (distinct from the sketch dimension m of Section 3.5). Draw $T_j \sim \nu/m_f$, $\omega_j \mid T_j \sim \mathcal{N}(0, 2T_j I_d)$, $\beta_j \sim \text{Unif}[0, 2\pi]$, and set $z(x) = (\sqrt{m_f/D} \sqrt{2} \cos(\omega_j^\top x + \beta_j) u(x))_{j=1}^D$. Then $\mathbb{E}[z(x)^\top z(w)] = k(x, w)$, and if $\|u(x)\| \leq B$ on X , then $\text{Var}[z(x)^\top z(w)] \leq 3m_f^2 B^4 / (2D)$ and, with probability $\geq 1 - \delta$, $\sup_{i,j} |z(x_i)^\top z(x_j) - k(x_i, x_j)| \leq m_f B^2 \sqrt{8 \log(2N^2/\delta)/D}$.*

We call these *Bernstein–Schur kernels*. The biased \mathbf{E} -kernel is the case $u = p_b$, $f(r) = (r + \varepsilon)^{-1}$ with $d\nu(t) = e^{-\varepsilon t} dt$, so the mass is $m_f = \nu(\mathbb{R}_{\geq 0}) = 1/\varepsilon$ and the law ν/m_f is $\text{Exp}(\varepsilon)$; with $B = \max_x \|p_b(x)\| = R^2 + b$ the variance scale $m_f^2 B^4 = (R^2 + b)^4 / \varepsilon^2$ recovers the bounds of Section 4. Theorem 3.6 establishes unbiasedness, the variance, and the uniform entrywise bound for the whole class; the matrix-Bernstein (Theorem 4.2) and kernel-ridge (Theorems 4.7 and 4.8) statements we prove for $k_{\mathbf{E}, b}$ and then extend to every member in Theorem 4.13, with P replaced by m_f times the bounded modulation Gram. The class is an elementary closure construction, but it is useful because it packages a family of kernels that are often neither stationary nor pure dot-product, and it permits unified unbiased features, exact variance formulas, and matrix-level approximation guarantees. Table 1 lists nontrivial members; only u and ν change between them.

Table 1: Bernstein–Schur kernels $k = p \cdot f$. Each row satisfies the finite-mass hypothesis $m_f = f(0) < \infty$ of Theorem 3.6 and is linearized by the same estimator: keep the finite feature of the modulation p exact, draw the radial scale from the (normalized) Bernstein measure of f , and apply random Fourier features to the Gaussian. The “mixing” column gives the normalized law ν/m_f (the IMQ has $d\nu = e^{-\varepsilon t} dt$, mass $m_f = 1/\varepsilon$; the generalized IMQ $(r + \varepsilon)^{-\alpha}$ has $d\nu \propto t^{\alpha-1} e^{-\varepsilon t} dt$, mass $m_f = \varepsilon^{-\alpha}$; the Matérn- $\frac{1}{2}$ radial $e^{-\sqrt{r}/\sigma}$, completely monotone in $r = \|x - w\|^2$, has the Lévy/inverse-Gaussian mixture with mass $m_f = 1$ and the exact two-line sampler $T = 1/(2\sigma^2 Z^2)$, $Z \sim \mathcal{N}(0, 1)$). Here $\Gamma(\alpha, \varepsilon)$ denotes shape α and rate ε . Only the two columns right of f change between instances.

Modulation $p(x, w)$	Radial $f(r)$	Mixing ν	feat. dim d_p	Inductive bias
$(x^\top w + b)^2, b \geq 0$	IMQ $(r + \varepsilon)^{-1}$	$\text{Exp}(\varepsilon)$	$\frac{d(d+1)}{2} + d + 1$	alignment \times proximity (biased \mathbf{E})
$(x^\top w + b)^q, q \in \mathbb{N}, b \geq 0$	$(r + \varepsilon)^{-\alpha}, \alpha > 0$ (gen. IMQ)	$\Gamma(\alpha, \varepsilon)$	$\binom{d+q}{q}$	q -way interactions \times locality
$(x^\top w + b)^q, q \in \mathbb{N}, b \geq 0$	Matérn- $\frac{1}{2}$ $e^{-\sqrt{r}/\sigma}$	$T = \frac{1}{2\sigma^2 Z^2}, Z \sim \mathcal{N}(0, 1)$ (Lévy)	$\binom{d+q}{q}$	q -way interactions \times exponential locality
$x^\top w + b, b \geq 0$	finite-mass compl. monotone f	Bernstein meas. of f	$d + 1$	signed local-linear trends
$u(x)^\top u(w)$	IMQ / Gaussian mix	corresponding	$\dim u$	data-driven modulation \times proximity

These other members are linearized by the same estimator in practice, not only in principle: on a genuinely non- \mathbf{E} instance (a degree-3 modulation $(x^\top w + b)^3$ times a generalized-IMQ radial $(\|x - w\|^2 + \varepsilon)^{-2}$ with $\Gamma(2, \varepsilon)$ mixing), the same scheme (keep u exact, draw $T_j \sim \Gamma(2, \varepsilon)$, apply RFF) is unbiased and converges at the Monte-Carlo rate, exactly as for the \mathbf{E} -kernel (Appendix D). The class-level theorem is thus more than a formal generalization: swapping u and ν in one estimator linearizes a different nonstationary kernel. The class also has standing customers outside this paper. Compositional kernel search (the Automatic Statistician, Duvenaud et al., 2013) emits GP kernels as products such as $\text{LIN}^2 \times \text{RQ}$, and the rational-quadratic radial $\text{RQ}_\alpha(r) = (1 + r/(2\alpha\sigma^2))^{-\alpha}$ is the generalized IMQ of Table 1 after rescaling ($\varepsilon' = 2\alpha\sigma^2$, mass $m_f = \text{RQ}_\alpha(0) = 1$, mixing law $\Gamma(\alpha, \varepsilon')$), so every such grammar product is a Bernstein–Schur kernel and inherits the estimator and the full matrix-level guarantee set with no new analysis, where previously it had no scaling story past the Gram. Run on California housing (off-sphere, $d = 8$, $\alpha = 2$), the class estimator of $(x^\top w + b)^2 \text{RQ}_\alpha$ converges to the exact composite Gram at the Monte-Carlo rate (fitted slope -0.47), matches exact composite-kernel ridge regression already at $D = 200$ (0.540 vs. 0.543 test RMSE), and fits the full $N = 19,640$ training set in 1.2 seconds as an explicit-feature primal where the $N \times N$ Gram has 4×10^8 entries (Appendix D).

3.5 Step 5: Sketching the modulation, the deployed estimator

Definition 3.3 keeps the modulation feature p_b exact, so its dimension $d_b = O(d^2)$ fixes the feature size $D D' d_b$. The deployed estimator removes this floor by randomizing the modulation as well. Let TS_m be a degree-2 TensorSketch of the augmented input (x, \sqrt{b}) , so $\mathbb{E}[\text{TS}_m(x)^\top \text{TS}_m(w)] = (x^\top w + b)^2$ over the sketch randomness (Pham & Pagh, 2013; Avron et al., 2014), and set $\hat{p}_m(x) = \varepsilon^{-1/2} \text{TS}_m(x) \in \mathbb{R}^m$. The deployed RAY map (randomizing both factors) replaces step 4 of Definition 3.3 by \hat{p}_m (using the flat $D' = 1$ form):

$$\hat{z}(x) = D^{-1/2} (\sqrt{2} \cos(\omega_j^\top x + \beta_j) \hat{p}_m(x))_{j=1}^D \in \mathbb{R}^{Dm}, \quad \omega_j \sim \mathcal{N}(0, 2t_j I_d), \quad t_j \sim \text{Exp}(\varepsilon). \quad (9)$$

Both factors are now random (the radial scale by the Bernstein mixture, the modulation by the sketch) and the feature dimension Dm is free of d^2 . The experiments (Section 5.2) and Proposition A.9 instead use a lower-variance *quadratic-only* variant that sketches only the degree-2 term $(x^\top w)^2$ and keeps the linear and constant terms $2b x^\top w + b^2$ exact, of dimension $D(m+d+1)$: still free of the $O(d^2)$ floor, larger only by an additive $O(d)$ per draw. Both forms are covered by Theorem 4.4 (the analysis touches the modulation only through \hat{P}_m). The exact-modulation estimator (8) (Definition 3.3) is the limit $m \rightarrow \infty$ ($\hat{p}_m \rightarrow p_b$), where the modulation randomness vanishes; that limit is the object the variance and concentration identities of Section 4 are stated for, and Theorem 4.4 carries the operator-norm guarantee to finite m at the cost of one additional, controllable sketch term. The modulation randomizer is modular: any unbiased or low-rank PSD modulation feature plugs in (TensorSketch, random-Maclaurin products, or anchor features), the choice being a geometry-dependent design decision rather than part of the analysis (Section 6). We state the guarantees for exact modulation and then transfer them, so the next two sections read as: clean identities at the limit, then the doubly-randomized bound.

Example 3.7 (One RAY coordinate, end to end). *Fix $b = 1$, $\varepsilon = 1$, $d = 3$. To build a single radial draw of the deployed map (9): (1) draw a scale $t \sim \text{Exp}(1)$, say $t = 0.4$; (2) draw a frequency $\omega \sim \mathcal{N}(0, 2tI_3)$ and phase $\beta \sim \text{Unif}[0, 2\pi)$, giving the radial Fourier feature $\sqrt{2} \cos(\omega^\top x + \beta)$, unbiased for the Gaussian $e^{-t\|x-w\|^2}$ at that scale; (3) form the modulation sketch $\hat{p}_m(x) = \text{TS}_m(x, \sqrt{b})$, unbiased for $(x^\top w + 1)^2$; (4) tensor them, $z_1(x) = \sqrt{2} \cos(\omega^\top x + \beta) \hat{p}_m(x) \in \mathbb{R}^m$. Then $\mathbb{E}[z_1(x)^\top z_1(w)] = e^{-t\|x-w\|^2} (x^\top w + 1)^2$, and averaging the scale over $t \sim \text{Exp}(1)$ recovers $h_\varepsilon \cdot p_b = k_{\mathbb{E},1}$. Stacking D such coordinates and dividing by \sqrt{D} gives the \mathbb{R}^{Dm} map. Setting $b = 0$ and replacing \hat{p}_m by the exact $\text{vec}(x \otimes x)$ recovers the unbiased exact-modulation feature of Definition 3.3.*

Remark 3.8 (Complex modulation sketches). *The modulation randomizer is the only part of RAY that need not be real. Wacker et al. (2024) show that complex-valued polynomial sketches have strictly smaller variance than real Rademacher sketches for dot-product kernels, because the unit-modulus fourth moment is smaller; substituting unit-phase signs $s \in \{1, i, -1, -i\}$ in the quadratic-only TS_m and reading off $\text{Re}\langle \text{TS}_m(x), \overline{\text{TS}_m(w)} \rangle$ leaves the estimator unbiased while lowering the additive sketch term in Theorem 4.4. The radial factor is unaffected. Measured on the protocol of Figure 2, the drop-in is real at every sketch size: the spectral error η falls by $1.5\text{--}1.7\times$ (at $m=128$, $0.042 \rightarrow 0.027$), the sketch term $\|E_P \circ R\|_{\text{op}}$ by the same factor ($3.84 \rightarrow 2.51$), and the per-entry sketch variance is halved (ratios $0.43\text{--}0.52$ across $m \in \{64, \dots, 512\}$; Appendix E, `complex_sketch`), consistent with the fourth-moment computation of Wacker et al. (2024).*

4 Approximation Guarantees

4.1 Variance and uniform error

Throughout, $X \subset \mathbb{R}^d$ is compact with $\|x\| \leq R$ for all $x \in X$. Proofs are in Appendix A.

The variance of the recommended flat estimator ($D' = 1$) is known in closed form. A coarser envelope valid for all D, D' ($V_D \leq \frac{(R^2+b)^4}{D\varepsilon^2} (1 + \frac{3}{2D'})$, the $3/(2D')$ term being the inner random-Fourier-feature variance) and the two-level identity (20) whose budget argument singles out $D' = 1$ as optimal are deferred to Appendix A (Theorem A.1, Proposition A.3); the experiments confirm $D' = 1$ directly (Appendix D.3).

Theorem 4.1 (Exact variance of the flat estimator). *Write $r = \|x - w\|^2$ and $a = (x^\top w + b)^2$. The flat estimator, one inner frequency per scale, $D' = 1$ (not the $D' \rightarrow \infty$ inner-averaged estimator), $\hat{k}_D =$*

$\frac{a}{\varepsilon} \frac{1}{D} \sum_{j=1}^D 2 \cos(\omega_j^\top x + \beta_j) \cos(\omega_j^\top w + \beta_j)$, with $T_j \sim \text{Exp}(\varepsilon)$ and $\omega_j \mid T_j \sim \mathcal{N}(0, 2T_j I_d)$, has

$$\text{Var}[\widehat{k}_D(x, w)] = \frac{a^2}{D\varepsilon^2} \left[1 + \frac{1}{2} \frac{\varepsilon}{\varepsilon + 4r} - \left(\frac{\varepsilon}{\varepsilon + r} \right)^2 \right].$$

At $r = 0$ the bracket is $\frac{1}{2}$ (the one-frequency variance does not vanish); the $D' \rightarrow \infty$ inner-averaged estimator instead has outer-only variance $\frac{a^2}{D\varepsilon^2} [\frac{\varepsilon}{\varepsilon+2r} - (\frac{\varepsilon}{\varepsilon+r})^2]$, the V_{out} term of Proposition A.3.

The identity locates the variance precisely: it is dominated by the fourth power of the bias-shifted alignment $a^2 = (x^\top w + b)^4$ and scales as ε^{-2} , while the radial bracket equals $\frac{1}{2}$ at $r = 0$, tends to 1 as $r \rightarrow \infty$, and is uniformly bounded by $\frac{3}{2}$. So the alignment numerator, not the distance, sets the scale; this recovers the $O((R^2 + b)^4 / (D\varepsilon^2))$ order of Theorem A.1 with exact constants.

A Hoeffding bound per entry plus a union bound gives a uniform entrywise Gram error $\sup_{ij} |z(x_i)^\top z(x_j) - k_{\mathbf{E},b}| \leq (R^2 + b)^2 \varepsilon^{-1} \sqrt{8 \log(2N^2/\delta)/D}$ with probability $1 - \delta$ (Theorem A.4), hence a sample complexity $D = O((R^2 + b)^4 \varepsilon^{-2} \tau^{-2} \log(N/\eta))$ free of explicit d (Corollary A.5); both are in Appendix A. Converting this entrywise bound to operator norm through $\|A\|_{\text{op}} \leq N \max_{ij} |A_{ij}|$ costs a factor N (Corollary A.6). That factor is wasteful: it ignores that the per-draw error matrices are structured. Each radial draw contributes $K^{(j)} = (\Psi_j \Psi_j^\top) \circ P$ with $P = [(x_i^\top x_j + b)^2 / \varepsilon]$, a Schur product of two positive semidefinite matrices and hence itself PSD; matrix Bernstein exploits this.

Theorem 4.2 (Expected matrix-Bernstein operator-norm bound). *Let $K = [k_{\mathbf{E},b}(x_i, x_j)]$ and $P = [(x_i^\top x_j + b)^2 / \varepsilon]$ be the kernel and polynomial Gram matrices (both PSD), and let K_D be the exact-modulation estimate from D i.i.d. radial draws. With $V = \sum_j \mathbb{E}[(D^{-1}(K^{(j)} - K))^2]$ the matrix variance and $d_{\text{int}} = \text{tr}(V) / \|V\|_{\text{op}} \leq N$ its intrinsic dimension (if $V = 0$ the estimate is exact and the bound holds trivially with the convention $d_{\text{int}} = 1$),*

$$\mathbb{E} \|K_D - K\|_{\text{op}} \leq 3 \sqrt{\frac{2\|P\|_{\text{op}} \|K\|_{\text{op}} \log(8d_{\text{int}})}{D}} + \frac{6\|P\|_{\text{op}} \log(8d_{\text{int}})}{D}.$$

The leading term scales with the top eigenvalues $\|P\|_{\text{op}}, \|K\|_{\text{op}}$ and the effective rank d_{int} , all $\ll N$ for spectrally concentrated data; since $\|P\|_{\text{op}} \leq \text{tr}(P) \leq N(R^2 + b)^2 / \varepsilon$ and $d_{\text{int}} \leq N$, the bound never exceeds the order of Corollary A.6 and is data-adaptively tighter. The theorem is stated for exact modulation; Theorem 4.13 extends it, with the rest of the matrix-level guarantee set, to every Bernstein–Schur kernel after replacing P by the bounded modulation Gram multiplied by the finite mass m_f . The key step is that $K^{(j)} \succeq 0$, so $(K^{(j)})^2 \preceq \|K^{(j)}\|_{\text{op}} K^{(j)} \preceq 2\|P\|_{\text{op}} K^{(j)}$ (the Schur-multiplier bound, Lemma A.7, gives $\|K^{(j)}\|_{\text{op}} \leq 2\|P\|_{\text{op}}$). Hence $\|V\|_{\text{op}} \leq 2\|P\|_{\text{op}} \|K\|_{\text{op}} / D$, and the intrinsic matrix-Bernstein inequality (Tropp, 2012; 2015) yields the bound. The full proof is in Appendix A; an empirical check across datasets of varying spectral spread confirms the data-adaptive constant (Appendix D.1).

Corollary 4.3 (High-probability operator-norm bound). *With the summands $Y_j = D^{-1}(K^{(j)} - K)$, the almost-sure bound $\|Y_j\|_{\text{op}} \leq L := 3\|P\|_{\text{op}} / D$, and the matrix variance $\|\sum_j \mathbb{E}[Y_j^2]\|_{\text{op}} \leq v := 2\|P\|_{\text{op}} \|K\|_{\text{op}} / D$, the tail form of the intrinsic matrix-Bernstein inequality (Tropp, 2015, Thm. 7.3.1) gives, with probability at least $1 - \delta$,*

$$\|K_D - K\|_{\text{op}} \leq 2 \sqrt{\frac{\|P\|_{\text{op}} \|K\|_{\text{op}} \log(8d_{\text{int}}/\delta)}{D}} + \frac{2\|P\|_{\text{op}} \log(8d_{\text{int}}/\delta)}{D}. \quad (10)$$

The proof (Appendix A) establishes L and v from $K^{(j)} \succeq 0$ and the Schur-multiplier bound (Lemma A.7), then inverts the intrinsic Bernstein tail.

The same conditioning argument carries this bound to the doubly-randomized estimator of Step 5, where the modulation is also sketched, the map one actually deploys.

Theorem 4.4 (Operator-norm error of RAY). *Let $\widehat{K}_{D,m} = \widehat{P}_m \circ \widehat{R}_D$ be the doubly-randomized estimator (9): a single degree-2 TensorSketch of dimension m (drawn once, shared across draws) gives the modulation Gram $\widehat{P}_m = [\text{TS}_m(x_i)^\top \text{TS}_m(x_j)] \succeq 0$ with $\mathbb{E} \widehat{P}_m = P = [(x_i^\top x_j + b)^2 / \varepsilon]$, and D radial draws give the radial Gram \widehat{R}_D with $\mathbb{E} \widehat{R}_D = R$, the unit-diagonal Gram $R_{ij} = \varepsilon / (\varepsilon + \|x_i - x_j\|^2)$ (so $K = R \circ P$), independent of the*

sketch. Suppose the sketch satisfies the spectral event $\|\widehat{P}_m - P\|_{\text{op}} \leq \eta\|P\|_{\text{op}}$ (Remark 4.5 discusses when the degree-2 TensorSketch achieves it). Then on that event, with probability at least $1 - \delta$ over the radial draws,

$$\|\widehat{K}_{D,m} - K\|_{\text{op}} \leq \underbrace{2\sqrt{\frac{(1+\eta)\|P\|_{\text{op}}\|K_S\|_{\text{op}}\log(8d_{\text{int},S}/\delta)}{D}} + \frac{2(1+\eta)\|P\|_{\text{op}}\log(8d_{\text{int},S}/\delta)}{D}}_{\text{radial, } O(D^{-1/2})} + \underbrace{\eta\|P\|_{\text{op}}}_{\text{sketch}}, \quad (11)$$

where $K_S = \widehat{P}_m \circ R$, $\|K_S\|_{\text{op}} \leq (1+\eta)\|P\|_{\text{op}}$, and $d_{\text{int},S} = \text{tr}(V_S)/\|V_S\|_{\text{op}} \leq N$ is the intrinsic dimension of the sketch-conditioned matrix variance V_S (the variance of Theorem 4.2 for the pair (\widehat{P}_m, K_S) rather than (P, K) , reducing to d_{int} as $m \rightarrow \infty$). At $m \rightarrow \infty$ ($\eta \rightarrow 0$) the sketch term vanishes and (11) is exactly Corollary 4.3: randomizing the modulation costs the single additive term $\eta\|P\|_{\text{op}}$, set by m independently of D .

Conditioning is the whole argument. Given the sketch, \widehat{P}_m is a fixed PSD modulation Gram and $\widehat{K}_{D,m} = \widehat{P}_m \circ \widehat{R}_D$ is an exact-modulation estimator of $K_S = \widehat{P}_m \circ R$; Theorem 4.2 and Corollary 4.3 apply verbatim with $P \mapsto \widehat{P}_m$ (so $\|\widehat{P}_m\|_{\text{op}} \leq (1+\eta)\|P\|_{\text{op}}$ and $\|K_S\|_{\text{op}} \leq \|\widehat{P}_m\|_{\text{op}}$ by Lemma A.7 with R unit-diagonal), giving the radial term. The remaining bias is $K_S - K = (\widehat{P}_m - P) \circ R = R \circ E_P$ with $E_P := \widehat{P}_m - P$ symmetric but *not* PSD; here the symmetric Schur-multiplier bound Lemma A.7(b) applies (with $A = R \succeq 0$, $\max_i R_{ii} = 1$), giving $\|R \circ E_P\|_{\text{op}} \leq \|E_P\|_{\text{op}} \leq \eta\|P\|_{\text{op}}$. The triangle inequality combines the two; the split is validated in Section 5.2. The conservative entrywise Frobenius decomposition of Proposition A.9 needs no subspace-embedding hypothesis and remains available as a fallback (Appendix A).

Remark 4.5 (Achieving the spectral event). *The degree-2 TensorSketch is an oblivious subspace embedding (Pham & Pagh, 2013; Avron et al., 2014): with sketch dimension m polynomial in the statistical dimension $s_\lambda = \text{tr}(P(P + \lambda I)^{-1})$ and in η^{-1} (the exact dependence is sketch- and degree-specific: up to degree-dependent factors, and the original degree-2 TensorSketch bound of Avron et al. (2014) is superlinear in s_λ), it delivers, with probability $1 - \delta_s$, a $(1 \pm \eta)$ ridge subspace embedding $(1 - \eta)(P + \lambda I) \preceq \widehat{P}_m + \lambda I \preceq (1 + \eta)(P + \lambda I)$. This is a regularized guarantee: the absolute event $\|\widehat{P}_m - P\|_{\text{op}} \leq \eta\|P\|_{\text{op}}$ assumed in Theorem 4.4 is its $\lambda \rightarrow 0$ idealization (there $s_\lambda \rightarrow \text{rank}(P)$ and the count becomes $\Omega(\eta^{-2} \text{rank}(P))$). The ridge-embedding event is the more useful one, because it transfers through the Schur product without loss: Proposition 4.6 carries it from the modulation Gram P to the kernel Gram K at the same ridge scale, so the deployed estimator inherits a ridge-relative sketch guarantee with sketch size polynomial in $s_\lambda(P)$ rather than $\text{rank}(P)$. The entrywise decomposition of Proposition A.9 needs no embedding hypothesis at all.*

Proposition 4.6 (Ridge-relative sketch transfer). *Suppose the modulation sketch is a $(1 \pm \eta)$ ridge subspace embedding of P at scale λ , $(1 - \eta)(P + \lambda I) \preceq \widehat{P}_m + \lambda I \preceq (1 + \eta)(P + \lambda I)$ (Remark 4.5). Then the exact-radial deployed Gram $K_S = \widehat{P}_m \circ R$ satisfies the same sandwich at the same scale,*

$$(1 - \eta)(K + \lambda I) \preceq K_S + \lambda I \preceq (1 + \eta)(K + \lambda I), \quad \text{equivalently} \quad \|A^{-1/2}(K_S - K)A^{-1/2}\|_{\text{op}} \leq \eta, \quad (12)$$

with $A = K + \lambda I$. Consequently $\|K_S - K\|_{\text{op}} \leq \eta(\|K\|_{\text{op}} + \lambda) \leq \eta(\|P\|_{\text{op}} + \lambda)$, sharpening the additive sketch term $\eta\|P\|_{\text{op}}$ of Theorem 4.4 whenever $\lambda \ll \|P\|_{\text{op}}$.

Proof. The radial Gram R is PSD with unit diagonal, so $R \circ I = I$ and hence $R \circ (M + \lambda I) = R \circ M + \lambda I$ for every symmetric M ; in particular $R \circ (P + \lambda I) = K + \lambda I$ and $R \circ (\widehat{P}_m + \lambda I) = K_S + \lambda I$. Both gaps in the hypothesis sandwich, $(1 + \eta)(P + \lambda I) - (\widehat{P}_m + \lambda I)$ and $(\widehat{P}_m + \lambda I) - (1 - \eta)(P + \lambda I)$, are PSD; Schur-multiplying each by $R \succeq 0$ keeps it PSD (Lemma A.7(a)), which is (12) after substituting the two identities. Conjugating by $A^{-1/2}$ gives the whitened form. For the last claim, (12) gives $-\eta(K + \lambda I) \preceq K_S - K \preceq \eta(K + \lambda I)$, so $\|K_S - K\|_{\text{op}} \leq \eta(\|K\|_{\text{op}} + \lambda)$, and $\|K\|_{\text{op}} \leq \|P\|_{\text{op}}$ by Lemma A.7 (R unit-diagonal). \square

This propagates to downstream kernel ridge regression: the approximation perturbs the learned predictor only through the operator-norm Gram error.

Theorem 4.7 (KRR stability under relative spectral approximation). *For a target $y \in \mathbb{R}^N$ and ridge $\lambda > 0$, let $\hat{\alpha} = (K + \lambda I)^{-1}y$, $\tilde{\alpha} = (K_D + \lambda I)^{-1}y$, $A = K + \lambda I$, and $E = K_D - K$. If the whitened error obeys*

$\|A^{-1/2}EA^{-1/2}\|_{\text{op}} \leq \rho < 1$, then

$$(1 - \rho)A \preceq K_D + \lambda I \preceq (1 + \rho)A, \quad \|\tilde{\alpha} - \hat{\alpha}\|_A \leq \frac{\rho}{1 - \rho} \|\hat{\alpha}\|_A,$$

where $\|v\|_A^2 = v^\top Av$.

This is the spectral-approximation form standard in random-feature KRR analysis (Avron et al., 2017): control of the *relative* error ρ , not the raw $\|K_D - K\|_{\text{op}}$, governs the predictor. The theorem is deterministic; Theorem 4.8 supplies the high-probability condition by running the matrix-Bernstein argument of Theorem 4.2 on the *whitened* summands $A^{-1/2}(K^{(j)} - K)A^{-1/2}$, where the ridge effective dimension $d_{\text{eff}}(\lambda) = \text{tr}(K(K + \lambda I)^{-1})$ appears as the intrinsic dimension of the whitened matrix variance. A cruder coefficient bound $\|\hat{\alpha} - \tilde{\alpha}\|_2 \leq \lambda^{-2} \|K_D - K\|_{\text{op}} \|y\|_2$ also holds, since $K_D = ZZ^\top \succeq 0$ makes $K_D + \lambda I$ invertible with inverse norm $\leq 1/\lambda$. It matches the $1/\sqrt{D}$ convergence observed downstream (Appendix D.1).

Theorem 4.8 (Whitened matrix Bernstein: the high-probability KRR condition). *Let $A = K + \lambda I$ with $\lambda > 0$, let K_D be the exact-modulation estimate from D i.i.d. radial draws, and set $\rho_D = \|A^{-1/2}(K_D - K)A^{-1/2}\|_{\text{op}}$. With the ridge effective dimension $d_{\text{eff}}(\lambda) = \text{tr}(KA^{-1})$, $\kappa_\lambda = \|K\|_{\text{op}}/(\|K\|_{\text{op}} + \lambda) < 1$, and $\tilde{d}_\lambda = d_{\text{eff}}(\lambda)/\kappa_\lambda$, with probability at least $1 - \delta$,*

$$\rho_D \leq 2\sqrt{\frac{\kappa_\lambda \|P\|_{\text{op}} \log(8\tilde{d}_\lambda/\delta)}{\lambda D}} + \frac{2}{3}\left(1 + \frac{2\|P\|_{\text{op}}}{\lambda}\right)\frac{\log(8\tilde{d}_\lambda/\delta)}{D}. \quad (13)$$

In particular, for any target $\rho_0 \in (0, 1]$,

$$D \geq \frac{16}{\rho_0^2}\left(1 + \frac{\|P\|_{\text{op}}}{\lambda}\right)\log\frac{8\tilde{d}_\lambda}{\delta} \implies \mathbb{P}\{\rho_D \leq \rho_0\} \geq 1 - \delta. \quad (14)$$

The proof (Appendix A) whitens the three ingredients of Theorem 4.2: the per-draw positivity $K^{(j)} \succeq 0$ survives conjugation, the a.s. bound becomes $(1 + 2\|P\|_{\text{op}}/\lambda)/D$, and the variance majorant becomes $\frac{2\|P\|_{\text{op}}}{\lambda D} A^{-1/2}KA^{-1/2}$, whose intrinsic dimension is *exactly* \tilde{d}_λ : the effective dimension is not an assumption imported from the KRR literature but the intrinsic dimension of the whitened variance. This closes the condition Theorem 4.7 left open. Numerically the identity is exact: across $\lambda \in \{10, 1, 0.1, 0.01\}$ the intrinsic dimension of $A^{-1/2}KA^{-1/2}$ matches \tilde{d}_λ to machine precision (relative error $\leq 4 \times 10^{-15}$), the whitened ρ_D decays at the $O(D^{-1/2})$ rate, the objective value stays in the Corollary 4.9 sandwich on every seed where $\rho_D < 1$, and the class instance of Theorem 4.13, the polynomially modulated Matérn- $\frac{1}{2}$ kernel with its Lévy sampler, reproduces all three (Appendix E).

Corollary 4.9 (High-probability KRR guarantee). *On the event $\rho_D \leq \rho_0 < 1$ (so, under the draw count (14), with probability at least $1 - \delta$), Theorem 4.7 applies: $(1 - \rho_0)A \preceq K_D + \lambda I \preceq (1 + \rho_0)A$ and $\|\tilde{\alpha} - \hat{\alpha}\|_A \leq \frac{\rho_0}{1 - \rho_0} \|\hat{\alpha}\|_A$. Moreover the optimal value of the kernel-ridge objective, $\min_f \sum_i (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 = \lambda y^\top (K + \lambda I)^{-1}y$, is preserved to relative constants:*

$$\lambda y^\top (K_D + \lambda I)^{-1}y \in \left[\frac{1}{1 + \rho_0}, \frac{1}{1 - \rho_0}\right] \cdot \lambda y^\top (K + \lambda I)^{-1}y.$$

At $\rho_0 = \frac{1}{2}$ the coefficient error is at most $\|\hat{\alpha}\|_A$ and the objective value is preserved within $[\frac{2}{3}, 2]$.

The count (14) carries the factor $\|P\|_{\text{op}}/\lambda$ because plain $\text{Exp}(\varepsilon)$ sampling pays the worst-case per-draw norm. The radial randomness is cheap to reweight, and the right tilt is computable in closed form: it is the trace of the whitened per-draw Gram, and its mean over the base law is *exactly* the effective dimension. Reweighting by it replaces $\|P\|_{\text{op}}/\lambda$ with $d_{\text{eff}}(\lambda)$ itself, the leverage-sampled count of the random-feature KRR literature (Avron et al., 2017; Li et al., 2021), here obtained on the Gram side with the same matrix-Bernstein proof.

Theorem 4.10 (Leverage-weighted radial sampling achieves the effective-dimension count). *Let $\theta = (t, \omega, \beta)$ denote one radial draw with base law π ($t \sim \text{Exp}(\varepsilon)$, $\omega \mid t \sim \mathcal{N}(0, 2tI_d)$, $\beta \sim \text{Unif}[0, 2\pi]$), per-draw Gram $K^{(\theta)} = (\psi_\theta \psi_\theta^\top) \circ P$ with $\psi_\theta[i] = \sqrt{2} \cos(\omega^\top x_i + \beta)$, and fix $\lambda > 0$, $A = K + \lambda I$. Define the whitened draw leverage*

$$\bar{d}_\lambda(\theta) := \text{tr}(A^{-1}K^{(\theta)}) = \psi_\theta^\top (A^{-1} \circ P) \psi_\theta \geq 0, \quad \mathbb{E}_\pi[\bar{d}_\lambda] = d_{\text{eff}}(\lambda).$$

Draw $\theta_1, \dots, \theta_D$ i.i.d. from the tilted law $d\pi_\lambda^* = (\bar{d}_\lambda/d_{\text{eff}}) d\pi$ and set $K_D^* = \frac{1}{D} \sum_j \frac{d_{\text{eff}}}{\bar{d}_\lambda(\theta_j)} K^{(\theta_j)}$. Then K_D^* is unbiased, and with probability at least $1 - \delta$, with $\ell = \log(8\bar{d}_\lambda/\delta)$,

$$\rho_D = \|A^{-1/2}(K_D^* - K)A^{-1/2}\|_{\text{op}} \leq \sqrt{\frac{2d_{\text{eff}}(\lambda)\kappa_\lambda\ell}{D}} + \frac{2}{3} \frac{(1 + d_{\text{eff}}(\lambda))\ell}{D}. \quad (15)$$

In particular, for any $\rho_0 \in (0, 1]$,

$$D \geq \frac{8}{\rho_0^2} (1 + d_{\text{eff}}(\lambda)) \log \frac{8\bar{d}_\lambda}{\delta} \quad \implies \quad \mathbb{P}\{\rho_D \leq \rho_0\} \geq 1 - \delta, \quad (16)$$

replacing the factor $1 + \|P\|_{\text{op}}/\lambda$ of Theorem 4.8 by $1 + d_{\text{eff}}(\lambda)$. Since $d_{\text{eff}}(\lambda) \leq N$ for every λ , the leverage count never exceeds $O(N \log(\bar{d}_\lambda/\delta))$, whereas the uniform count grows without bound as $\lambda \rightarrow 0$.

The proof (Appendix A) whitens the same three ingredients with the weights in place: on the support of π_λ^* the weighted whitened draw has operator norm at most its trace, which the tilt normalizes to exactly d_{eff} ; the variance majorant becomes $\frac{d_{\text{eff}}}{D} A^{-1/2} K A^{-1/2}$, the same core matrix as in Theorem 4.8, so its intrinsic dimension is again exactly \bar{d}_λ . Two readings. (i) *What is being tilted.* The leverage $\bar{d}_\lambda(\theta)$ is a quadratic form in ψ_θ with the fixed PSD matrix $A^{-1} \circ P$: draws whose cosine pattern injects energy where A^{-1} is large (the small-eigenvalue directions of K that the ridge does not protect) are over-sampled and down-weighted. (ii) *Deployment.* Exact tilting needs $A^{-1} \circ P$, the usual leverage chicken-and-egg; the standard answers (a pilot uniform estimate, Nyström or sketched approximations of A^{-1} , iterative reweighting) apply verbatim since the proof only uses the importance identity and the trace normalization. Numerically the prediction lands as stated: on the protocol of Theorem 4.8 ($N=300$, $\|P\|_{\text{op}}=306$), the uniform draw count $D^*(\rho_D \leq \frac{1}{2})$ climbs $50 \rightarrow 800 \rightarrow >3200$ as λ drops $10 \rightarrow 1 \rightarrow 0.1$, while the leverage-tilted sampler reaches it at $D^* = 50/200/400/1600$ across $\lambda \in \{10, 1, 0.1, 0.01\}$, tracking $d_{\text{eff}} = 12.3/38.6/90.8/163.3$; the identity $\mathbb{E}_\pi[\bar{d}_\lambda] = d_{\text{eff}}$ holds on a 5×10^4 -draw pool to within 2%, and both samplers keep the $O(D^{-1/2})$ whitened rate (fitted slopes -0.47 to -0.63 ; Appendix E, `leverage_radial_sampling`).

Remark 4.11 (From spectral approximation to risk, and what leverage buys). *Three reads of Theorems 4.8 and 4.10.* (i) The counts. $D = O((1 + \|P\|_{\text{op}}/\lambda) \log d_{\text{eff}})$ is the plain-i.i.d.-sampling rate, the Gram-side analogue of the $\Omega(1/\lambda)$ feature counts for uniformly sampled random Fourier features in KRR (Avron et al., 2017; Rudi & Rosasco, 2017; Bach, 2017); Theorem 4.10 improves the leading factor to $d_{\text{eff}}(\lambda)$ itself, the ridge-leverage count of Avron et al. (2017); Li et al. (2021), by tilting the joint radial draw (t, ω, β) rather than the scale alone. What remains open on this axis is purely computational: a deployment-grade approximation of the tilt \bar{d}_λ (pilot estimates, Nyström approximations of A^{-1}) with the approximation error folded into the bound. (ii) Rates. The condition $\rho \leq \frac{1}{2}$ at the regularization schedule $\lambda = \lambda_N$ of the source/capacity conditions is the interface assumption of the random-feature KRR literature, and both counts stay polynomial in N along any such schedule, the leverage count uniformly so ($d_{\text{eff}} \leq N$). (iii) An honest caveat. A spectral sandwich alone does not bound the fixed-design risk: $K = 0$ with $K' = \rho\lambda I$ satisfies the ρ -sandwich yet inflates the in-sample variance term from 0 to $\sigma^2 \rho^2 / (1 + \rho)^2$. A risk theorem therefore needs, in addition, control of the approximate kernel's effective dimension $d_{\text{eff}}(K_D^*)$, which a plain sandwich does not supply; with the count side now closed, this is the one missing ingredient between Theorem 4.10 and a minimax-optimal RAY-KRR statement. The coefficient and objective-value guarantees of Corollary 4.9 hold for K_D^* as they stand.

The same conditioning device that carried Corollary 4.3 to the deployed estimator (Theorem 4.4) carries Theorem 4.8 as well.

Corollary 4.12 (Ridge-relative KRR condition for the deployed estimator). *Let $\widehat{K}_{D,m}$ be the RAY estimate (9) and fix a target $\rho_0 \in (0, 1]$. Suppose the sketch is a $(1 \pm \eta)$ ridge embedding of P at scale λ (Remark 4.5) with $\eta \leq \rho_0/4$. Then, on the sketch event, with probability at least $1 - \delta$ over the radial draws,*

$$D \geq \frac{64}{\rho_0^2} \left(1 + \frac{(1 + \eta)\|P\|_{\text{op}}}{\lambda}\right) \log \frac{8\bar{d}_{\lambda,S}}{\delta} \quad \implies \quad \|A^{-1/2}(\widehat{K}_{D,m} - K)A^{-1/2}\|_{\text{op}} \leq \rho_0,$$

and Corollary 4.9 holds for the deployed Gram $\widehat{K}_{D,m}$ verbatim. Here $\bar{d}_{\lambda,S}$ is the intrinsic dimension of the sketch-conditioned whitened variance.

Proof. Proposition 4.6 gives, on the sketch event, the deterministic bound $\rho_{\text{sk}} := \|A^{-1/2}(K_S - K)A^{-1/2}\|_{\text{op}} \leq \eta$ for the exact-radial Gram $K_S = \hat{P}_m \circ R$. Conditioned on the sketch, $\hat{K}_{D,m}$ is an exact-modulation estimate of K_S , so Theorem 4.8 applied to the pair (\hat{P}_m, K_S) with $\|\hat{P}_m\|_{\text{op}} \leq (1 + \eta)\|P\|_{\text{op}}$ and target $\rho_0/2$, whitened by $A_S = K_S + \lambda I$, gives $\|A_S^{-1/2}(\hat{K}_{D,m} - K_S)A_S^{-1/2}\|_{\text{op}} \leq \rho_0/2$ under the stated count. Since $A_S \preceq (1 + \eta)A$, the A -whitened norm is at most $(1 + \eta)$ times the A_S -whitened one, and the triangle inequality gives $\rho_0 \geq (1 + \eta)\frac{\rho_0}{2} + \eta$ for $\eta \leq \rho_0/4$. \square

The sketch requirement is now scale-free, $\eta \leq \rho_0/4$, a constant rather than the $\eta \lesssim \lambda/\|P\|_{\text{op}}$ of the absolute-event route, and the size is polynomial in the statistical dimension $s_\lambda(P)$ rather than $\text{rank}(P)$ (Remark 4.5). The radial count carries $\log \tilde{d}_{\lambda,S}$ rather than the worst-case $\log N$: the ridge embedding gives $\tilde{d}_{\lambda,S} \leq (\tilde{d}_\lambda + \eta N \kappa_\lambda^{-1})/(1 - \eta)$, so the count is the genuine effective-dimension count $\log \tilde{d}_\lambda$ once $\eta \lesssim \tilde{d}_\lambda/N$, degrading to $\log N$ only as the sketch coarsens. This is the honest content of the transfer: the *sketch accuracy* requirement drops to a constant unconditionally, while the *logarithmic dimension* is intrinsic only when the sketch is fine enough.

None of the matrix-level analysis is special to the \mathbf{E} -kernel. The proofs consume exactly three structural facts (a PSD per-draw Gram that is a Schur product with a bounded-diagonal rank-one factor, a PSD modulation Gram, and a unit-diagonal PSD radial Gram), and every Bernstein–Schur kernel supplies all three.

Theorem 4.13 (Class-level matrix concentration and KRR condition). *Let $k = p \cdot f$ be a Bernstein–Schur kernel in the setting of Theorem 3.6 (modulation feature u , mixing measure ν of finite mass m_f), let $G_u = [u(x_i)^\top u(x_j)]$ be the modulation Gram, and set $P_u := m_f G_u$. Then Theorem 4.2, Corollary 4.3, Theorem 4.8, Theorem 4.10, and Corollary 4.9 hold verbatim for the class estimator of Theorem 3.6 with P replaced by P_u . For $k_{\mathbf{E},b}$, $P_u = P$.*

The proof (Appendix C) verifies the three facts and nothing else. The statement upgrades the class from “unbiased with entrywise control” (Theorem 3.6) to the full operator-norm and kernel-ridge guarantee set; in particular the polynomially modulated Matérn- $\frac{1}{2}$ kernel $(x^\top w + b)^q e^{-\|x-w\|/\sigma}$, whose radial scale has the exact two-line sampler of Table 1, now carries them all.

Read across b , these bounds expose what the bias costs. Setting $b = 0$ recovers the unbiased bounds with R^4 and R^8 ; for $b > 0$ they inflate by powers of $(1 + b/R^2)$, the bias enlarging the effective data radius from R^2 to $R^2 + b$. This is the price side of the trade whose benefit, the finite-difference IMQ reach and the practical bias shift, motivated $k_{\mathbf{E},b}$ over $k_{\mathbf{E}}$ in the first place (Section 1); the $(R^2 + b)^4$ variance law is not merely an artifact of the proof but is borne out empirically in Appendix D.1.

The $(R^2 + b)^4$ blow-up is removable by normalizing the modulation feature to $q_b(x) = p_b(x)/(\|x\|^2 + b)$ (unit norm): the resulting *normalized estimator* has variance bounded by $3/(2D\varepsilon^2)$, free of R and b (Proposition A.8, Appendix A). It is not an approximation to $k_{\mathbf{E},b}$ but an unbiased estimator of a *different*, cosine-rescaled kernel, so the gain is stability, not a free fix.

Finally, the Gram error has a structure that explains why the product is worth keeping even when the modulation is itself compressed. Writing $K = P \circ H$ with $P_{ij} = p_b(x_i, x_j)$, $H_{ij} = h_\varepsilon(x_i, x_j)$:

Proposition 4.14 (Modulation–radial error decomposition). *With \hat{P}_m any modulation approximation and \hat{H}_D the radial estimate, $\hat{K}_{D,m} = \hat{P}_m \circ \hat{H}_D$ satisfies*

$$\hat{K}_{D,m} - K = \underbrace{P \circ (\hat{H}_D - H)}_{\text{radial error, modulated by finite feature}} + \underbrace{(\hat{P}_m - P) \circ H}_{\text{modulation error, localized by proximity}} + \underbrace{(\hat{P}_m - P) \circ (\hat{H}_D - H)}_{\text{interaction}}.$$

For exact modulation ($\hat{P}_m = P$) this collapses to $\hat{K}_D - K = P \circ (\hat{H}_D - H)$, so entrywise $|\hat{K}_{ij} - K_{ij}| = p_b(x_i, x_j) |\hat{H}_{ij} - H_{ij}|$.

(Proof: expand $(P + E_P) \circ (H + E_H)$ and subtract K .) The radial Monte-Carlo noise is Schur-scaled by the modulation; under bounded norms, and literally for the normalized variant where $G_{ij} = (x_i^\top x_j + b)^2 / ((\|x_i\|^2 + b)(\|x_j\|^2 + b)) \in [0, 1]$, this acts as an alignment gate, since p_b otherwise also carries norm and bias scale. Appendix D.1 shows this modulation suppresses false positives even when the polynomial factor is sketched.

4.2 Where the dimension enters, and what to compare against

Table 2 states the sample complexity for a τ -approximation of the N -point Gram matrix. We are careful about the setting, because it is easy to overclaim. At this *dataset level* (a union bound over the $\leq N^2$ pairs, via Hoeffding), the RAY radial count is free of d (Corollary A.5). But so is standard RFF: a union bound over a finite point set removes the explicit d for the Gaussian and IMQ kernels as well. Dimension-freeness of the sample count is therefore a property of the dataset-level analysis, not a special feature of our construction.

Table 2: Sample complexity for a dataset-level τ -approximation of the N -point Gram matrix (Hoeffding + union bound). All counts are free of the input dimension d at this level; the constant is the squared per-draw range. Standard RFF uses an absolutely bounded cosine feature, so its count is free of R and the bandwidth σ ; the IMQ–Laplace estimator carries the radial mass $m_f = 1/\varepsilon$, giving ε^{-2} . RAY’s count is exactly that radial ε^{-2} times the *unbounded* polynomial modulation range $(R^2 + b)^4$, the honest price of the alignment numerator, removed only by normalizing the modulation (Proposition A.8). The methods differ *qualitatively* in applicability, since RFF needs shift-invariance and sketching a dot-product form, and $k_{\mathbb{E},b}$ has neither. R is the data radius.

Kernel (structure)	Method	D for τ -approx
Gaussian (shift-invariant)	RFF	$O(\tau^{-2} \log N)$
IMQ h_ε (shift-invariant)	RFF–Laplace	$O(\varepsilon^{-2} \tau^{-2} \log N)$
$k_{\mathbb{E},b}$ (neither shift-inv. nor dot-product)	RAY (ours)	$O(\frac{(R^2+b)^4}{\varepsilon^2} \tau^{-2} \log N)$

Table 3: Approximation taxonomy. Random-feature methods draw feature parameters from kernel-defined distributions before seeing the dataset and are unbiased for arbitrary pairs. Nyström methods are data-dependent deployment baselines: they choose columns, centers, or leverage-sampled landmarks from the observed training set and approximate the empirical Gram matrix or downstream predictor.

Method	Data-dependent?	Unbiased for $k(x, w)$?	Streaming primal?
Exact modulation (analyzable limit)	No	Yes	Yes
Gaussian / IMQ RFF	No	Yes	Yes
RAY (sketched modulation)	No (sketch-random)	Yes, over all draws	Yes
Uniform Nyström	Yes (columns from X)	No	No
k-means Nyström	Yes (clustered X)	No	No
Ridge-leverage Nyström	Yes (scores from X)	No	No

What RAY genuinely buys is therefore not a smaller d -dependence than RFF, but *applicability*: it brings a kernel that is neither shift-invariant nor a dot-product, on which RFF and polynomial sketching cannot be run, into the same dataset-level, d -free regime that RFF enjoys for the kernels it does cover, at the cost of a larger constant: $(R^2 + b)^4/\varepsilon^2$ in place of the absolute constant of a bounded-cosine RFF, the unbounded polynomial modulation being the one factor RFF does not carry. The explicit d familiar from RFF reappears only in the *domain-level* (uniform-over-a- d -ball) analysis through the covering number; there RAY’s inner Gaussian features incur it as well, so RAY is not d -free in that stronger sense either. Nyström occupies a different category (Table 3): it is an adaptive landmark approximation to the observed sample, not a data-independent random-feature map. Its error can degrade with dimension through the spectral-decay rate $O(m^{-2s/d})$, which is the contrast our matched-landmark experiments target (Section D.1, Table 11), but its adaptivity can also make it a stronger deployment baseline at fixed representation size (Appendix D.1).

4.3 Variance reduction

Three optional refinements of the sampling lower the variance further; all are evaluated in Appendix D.

Quasi-Monte Carlo over t . Replace $t_j \sim \text{Exp}(\varepsilon)$ by quasi-random samples through the inverse CDF $t_j = -\varepsilon^{-1} \log u_j$, u_j from a Sobol or Halton sequence. The radial integrand $t \mapsto e^{-t(\varepsilon + \|x-w\|^2)}$ is C^∞

and of bounded variation, so by Koksma–Hlawka (Kuo & Nuyens, 2016) the t -integral’s QMC error is $O((\log D)^s/D)$. We emphasize, however, that this is the rate of the outer integral only: the estimator also carries the inner Gaussian RFF Monte-Carlo noise, which is not quasi-randomized, and empirically (Appendix D.2) it is this inner noise that sets the overall convergence, so QMC delivers a constant-factor variance reduction rather than a faster rate. Deterministic Sobol/Halton points also trade the exact i.i.d. unbiasedness of Theorem 3.4 for a bounded deterministic integration error; *randomized* QMC (a scrambled or shifted sequence) restores unbiasedness in expectation over the randomization while keeping the improved error, and is the variant we recommend when unbiasedness must hold exactly.

Orthogonal features for the inner approximation. For each scale t_j , drawing the inner frequencies as scaled rows of a Haar-random orthogonal matrix (Yu et al., 2016) reduces the inner-loop variance without affecting unbiasedness. The combined QMC-outer / orthogonal-inner estimator achieves the best of both. At the recommended $D' = 1$ there is no inner block to decorrelate, and the right target is instead the *marginal* spectral measure of the radial factor (Section 3.4): drawing the D frequencies ω_j jointly orthogonal or quasi-random against that single isotropic distribution, with radii coupled to the inverse radial CDF, decorrelates the draws the current i.i.d. estimator leaves independent. This is free variance reduction and the natural form of the refinement when $D' = 1$.

Importance sampling over t . Sampling $t \sim \text{Exp}(\varepsilon + \eta)$ with weight $w_t = \frac{\varepsilon}{\varepsilon + \eta} e^{\eta t}$ keeps the estimator unbiased and reduces variance for nearby pairs ($\|x - w\|^2 \leq \eta$) by up to $(\varepsilon/(\varepsilon + \eta))^2$, *provided η is not too large*: the importance estimate of the radial factor at squared distance r has a finite second moment only for $\eta < \varepsilon + 2r$ (Appendix D), so a single proposal safe for every pair, including $r = 0$, requires $\eta < \varepsilon$. The proposal is data-independent and so practical within that range, if suboptimal for the full pairwise-distance spread.

Deterministic radial quadrature. The outer scale integral can be discretized deterministically rather than by Monte Carlo, replacing the across-scale variance of Proposition A.3 by a uniform, controllable bias. This answers the kernel-quadrature question of Section 6 for the radial factor in the affirmative.

Proposition 4.15 (Positive-weight radial quadrature). *The radial factor is the Laplace transform $h_\varepsilon(r) = \int_0^\infty e^{-\varepsilon t} e^{-tr} dt = 1/(\varepsilon + r)$. For every tolerance $\tau > 0$ there exist, by exponential-sum approximation of $1/x$ on a bounded interval (Beylkin & Monzón, 2010), $D = O(\log(1 + 4R^2/\varepsilon) \log(1/\tau))$ positive nodes $t_1, \dots, t_D > 0$ and positive weights $w_1, \dots, w_D > 0$ with*

$$\sup_{r \in [0, 4R^2]} \left| \sum_{j=1}^D w_j e^{-t_j r} - h_\varepsilon(r) \right| \leq \tau. \quad (17)$$

Replacing the random scales $t_j \sim \text{Exp}(\varepsilon)$ by these nodes and the average by $\sum_j w_j(\cdot)$ leaves a surrogate radial Gram that is a nonnegative combination of Gaussian kernels, so the surrogate $\hat{k} = (\sum_j w_j e^{-t_j \|\cdot - \cdot\|^2}) \cdot p$ is positive definite, and the entrywise error is the deterministic bias $\sup_{i,j} |\hat{k}_{ij} - k_{ij}| \leq \tau(R^2 + b)^2$, free of the $O(D^{-1/2})$ radial fluctuation of Theorem A.4.

Proof. Equation (17) is the cited exponential-sum bound applied to $1/x$ on $[\varepsilon, \varepsilon + 4R^2]$ after the shift $x = \varepsilon + r$; positive weights keep each $e^{-t_j \|x - w\|^2}$ a Gaussian kernel, so the nonnegative combination and its Schur product with the PSD modulation p are PSD. Multiplying (17) by the modulation, bounded by $(R^2 + b)^2$, gives the entrywise bias. \square

Each fixed-scale Gaussian still needs an inner feature, so the deployed dimension is unchanged; what the nodes remove is the outer scale sampling. The empirical scoping is sharp (Appendix E, `radial_quadrature`): with the Gaussian factor computed *exactly* (the Gram model), a D -node Gauss–Laguerre rule reaches machine precision by $D = 32$ where Monte-Carlo is still at 5.4×10^{-2} RMSE; but in the deployed estimator the inner Fourier noise dominates (Appendix D.2), and a naive equal-allocation node rule is in fact *worse* than i.i.d. sampling at matched draws (0.38 vs. 0.17 at $D = 32$), because the nodes spend budget on scales the random law visits in proportion to their weight. The nodes are therefore the right tool for exact-Gaussian surrogates

and small-Gram settings, and as the bounded scale sets that positive features require (Proposition D.2, through the largest node), not a drop-in replacement for the trigonometric sampler. The count grows only *logarithmically* with the radial dynamic range $4R^2/\varepsilon$; the price is the loss of exact unbiasedness, traded for the uniform bias (17), and the unbiased sampler remains the variant for the streaming and analysis claims.

5 Experiments

The experiments answer four questions in order. First, does RAY approximate the off-sphere kernel in the regime where neither shift-invariant RFF nor dot-product sketches apply? Second, does TensorSketch remove the $O(d^2)$ modulation floor while preserving the predicted radial-plus-sketch error split? Third, is the alignment \times proximity kernel useful exactly when the target couples those two factors? Finally, does the resulting feature map run in the streaming, Gram-impossible regime it is designed for? Cost-matched caveats, gating diagnostics, attention fidelity, sphere-normalized sanity checks, and other ablations are deferred to Appendix D. This keeps the main section focused on the empirical spine; the appendix records the negative and scoped results needed to interpret it honestly.

5.1 Off-sphere validation on a bounded ball

The key test is off-sphere (Figure 1): a bounded ball where $k_{\mathbf{E},b}$ is not a function of $x^\top w$ alone, so no dot-product reduction exists. We sample $N = 1000$ points with directions uniform on \mathbb{S}^{d-1} and radii uniform on $[0.25, 1]$ (a dimension-independent spread of norms; uniform-in-ball sampling would collapse toward the sphere as d grows). Approximating the exact Gram with RAY (flat $D' = 1$, $b = 1$, ε the median squared distance, 5 seeds) reproduces the $O(1/\sqrt{D})$ rate, while uniform- and k-means-landmark Nyström degrade with d (Table 4): k-means helps by a constant but not the trend, and RAY at $D = 1000$ is below both by $d = 8$. This matches radial draws against landmarks, not feature dimension (RAY’s is $D d_b$); Appendix D.1 gives the cost-matched comparison. Bounded norm, not sphere normalization, is what the estimator requires.

Table 4: Relative Frobenius Gram error on an *off-sphere* bounded ball ($\|x\| \in [0.25, 1]$, varying norms; $N = 1000$, $b = 1$, mean over 5 seeds). RAY follows the $O(1/\sqrt{D})$ rate at every dimension; both uniform and k-means Nyström ($m = 100$) worsen as d grows, k-means by a smaller constant. Standard deviation over the 5 seeds is ≤ 0.015 for the RAY columns and ≤ 0.003 for the Nyström columns, so the $d \geq 8$ ordering (RAY below both Nyström variants) is outside seed noise.

d	RAY, radial samples D			uniform Nyström	k-means Nyström
	10	100	1000	$m = 100$	$m = 100$
2	0.36	0.09	0.040	0.001	0.000
8	0.55	0.14	0.044	0.073	0.051
16	0.48	0.17	0.051	0.092	0.077
32	0.53	0.17	0.057	0.103	0.093

The adaptive *ridge-leverage-score* Nyström (Musco & Musco, 2017), the strongest standard baseline, does not change the picture: it tracks uniform Nyström (0.072/0.094/0.110 at $d = 8/16/32$), k-means stays the best landmark variant (0.051/0.079/0.097), and all three worsen with d while RAY at $D = 1000$ remains below them by $d = 16$. The behavior persists into genuinely high dimension: on the same off-sphere ball at $d \in \{128, 256, 512\}$, RAY follows the $O(1/\sqrt{D})$ rate (fitted slopes $-0.41, -0.51, -0.52$) with relative error 0.05–0.06 at $D = 1000$, unchanged in order across a $4\times$ range of d .

The fidelity carries to real data kept off-sphere. Standardizing and scaling each dataset by its maximum row norm (so $\|x\| \leq 1$ with *varying* norms, not on the sphere), RAY at $D = 128$ tracks the exact \mathbf{E} -kernel’s downstream accuracy on `digits` (0.983 vs. exact 0.986) and on a larger `covtype` subsample ($N=3000$, $d=54$: 0.695 vs. exact 0.682), in both cases matching or beating Gaussian RFF and k-means Nyström at the same budget. So the sphere-normalized ridge-regression results (Appendix D) are not an artifact of normalization: RAY reproduces the exact kernel off-sphere as well.

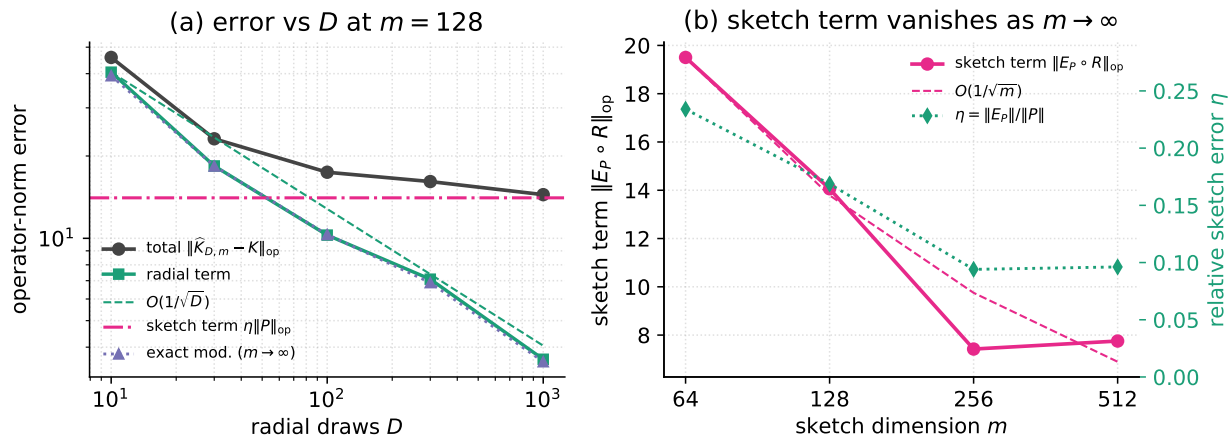


Figure 2: Operator-norm error of the deployed (doubly-randomized) RAY estimator, validating Theorem 4.4 (off-sphere, $d = 16$, $N = 300$, $\|P\|_{\text{op}} = 186$). (a) At fixed sketch size $m=128$, the radial term falls as $O(1/\sqrt{D})$ while the sketch term $\eta\|P\|_{\text{op}}$ is a D -independent floor; the total decays to that floor, and the $m \rightarrow \infty$ (exact-modulation) curve is the zero-floor limit. (b) The sketch term $\|E_P \circ R\|_{\text{op}}$ and the relative sketch error $\eta = \|E_P\|_{\text{op}}/\|P\|_{\text{op}}$ both shrink with m , so increasing m recovers exact modulation (Corollary 4.3). The single additive sketch term is the entire price of randomizing the modulation.

5.2 Compressing the polynomial factor with TensorSketch

RAY sketches the degree-2 modulation to escape the $O(d^2)$ feature floor (Section 3.5); here we validate the error decomposition and exercise the accuracy–cost trade. The variance splits into a radial Monte-Carlo term ($\sim D^{-1/2}$), a polynomial-sketch term ($\sim m^{-1/2}$), and their product (Proposition A.9). For a fixed pair (4000 repetitions) the split is clean: varying D at $m = 256$ drives only the radial term ($0.09 \rightarrow 8.5 \times 10^{-4}$), varying m at $D = 1000$ only the sketch term ($4.2 \times 10^{-3} \rightarrow 1.4 \times 10^{-4}$), and the empirical variance matches the three-term formula throughout (ratio 0.94–1.06). The operator-norm error of Theorem 4.4 shows the same structure (Figure 2): at $m=128$ the radial term falls as $D^{-1/2}$ ($40.5 \rightarrow 3.6$ over $D=10 \rightarrow 1000$) while the sketch term is a D -independent floor that decays with m ($19.5 \rightarrow 7.4$ as $m=64 \rightarrow 256$), vanishing as $m \rightarrow \infty$.

At finite m the compression recovers most of the efficiency the exact feature loses to its $O(d^2)$ size: on sphere-normalized digits ($d = 64$) at $M = d_b$, RAY ($m = 128$) reaches 0.977 where exact modulation, starved to a single radial draw, scores 0.928, near the optimal rank- M ceiling 0.986 (sanity check, Table 22, Appendix D). In low dimension there is nothing to compress (california, $d_b = 45$): the sketch only adds noise and one takes $m \rightarrow \infty$. RAY is the default scalable map; exact modulation is its low- d endpoint, not a separate method.

At a fixed feature dimension $M = D(m+d+1)$ the sketch size m trades against the radial-draw count D , and the optimum is interior (Table 5), at the closed-form $m^* = \sqrt{BM/A}$ derived from the two-term variance (Proposition A.10). The best allocation for Gram fidelity is an intermediate sketch size; downstream KRR error instead favors smaller m (more radial draws), so the operating point depends on whether Gram fidelity or prediction is the goal. This operationalizes the radial-vs-sketch split of Theorem 4.4.

Sketching also removes the $O(d^2)$ representation floor outright (Table 6). At $d=1024$ the exact-modulation feature needs 33.7 GB per the $N=1000$ representation and cannot be built, while the sketched feature is 74 MB and builds in 28 ms, a direct demonstration that sketching removes the $O(d^2)$ floor (Limitation (v)).

5.3 Coupled-target preference

The comparisons so far measure approximation quality, not whether the \mathbf{E} -kernel is the right kernel. We test that inductive bias directly: the kernel couples alignment and proximity, so it should be preferred exactly when the target benefits from both, and not otherwise. On off-sphere data ($\|x\| \in [0.3, 1.5]$, $d = 16$) we build

Table 5: Deployed-RAY allocation at fixed feature dimension $M = 8192$ on an off-sphere ball (relative Frobenius Gram error, mean over 3 seeds): at fixed $M = D(m+d+1)$ the sketch size m trades against the radial-draw count D . The error is minimized at an intermediate m (small m leaves a large sketch term, large m starves the radial draws), and the optimal m grows slowly with d .

m	$d=16$	$d=64$	$d=256$
16	0.212	0.252	0.424
32	0.173	0.216	0.397
64	0.173	0.216	0.305
128	0.185	0.228	0.376
256	0.240	0.281	0.343
512	0.520	0.488	0.423

Table 6: Feature-construction cost vs. dimension d at fixed radial draws $D=8$, sketch $m=128$, $N=1000$ (off-sphere ball). Exact modulation’s per-point feature dimension $D d_b$ grows as $O(d^2)$ and becomes impossible to build; the sketched feature $D(m+d+1)$ stays linear in d . “–” marks a feature too large to materialize ($> 2\text{M}$ coordinates/point).

d	exact dim	exact mem (GB)	exact build (s)	sketch dim	sketch mem (MB)	sketch build (s)
64	17160	0.14	0.043	1544	12.4	0.005
128	67080	0.54	0.143	2056	16.4	0.007
256	265224	2.12	1.585	3080	24.6	0.011
512	1054728	8.44	8.202	5128	41.0	0.017
1024	4206600	33.65	–	9224	73.8	0.028

three regression targets from small atom sums and compare kernel ridge regression with the Gaussian, IMQ, degree-2 polynomial, and \mathbf{E} kernels plus RAY (Table 7). The two single-factor controls are kernel-natural (*proximity* $y = \sum_k a_k / (\|x - v_k\|^2 + \varepsilon_0)$) and *alignment* $y = \sum_k a_k (u_k^\top x)^2$), so they should, and do, favor the distance kernels and the polynomial kernel respectively. The *coupled* target $y = \sum_k a_k \tanh(2u_k^\top x) e^{-\|x - v_k\|}$ is deliberately *not* of the \mathbf{E} -kernel’s form: it multiplies a tanh alignment by a Laplace proximity, matching no candidate kernel, so a \mathbf{E} win there cannot be an artifact of \mathbf{E} -generated data. The \mathbf{E} -kernel wins the coupled target and only there. The evidence is therefore a controlled compatibility result for the alignment \times proximity coupling, not a general necessity theorem. RAY inherits the preference once its approximation has converged: at $D = 4000$ it reaches 0.093 on the coupled target, below both distance kernels, approaching the exact \mathbf{E} -kernel’s 0.087; at smaller D the residual approximation error keeps it above them, consistent with the $O(1/\sqrt{D})$ convergence measured elsewhere.

Table 7: Kernel ridge regression test RMSE (\downarrow) on off-sphere targets ($d = 16$, varying norms, mean over 5 seeds, RAY at $D = 4000$, best per row in bold). The coupled target (tanh alignment \times Laplace proximity) matches no candidate kernel; the \mathbf{E} -kernel is best only there, while distance kernels win the proximity control and the polynomial wins the alignment control. RAY tracks the exact \mathbf{E} -kernel, beating both distance kernels on the coupled target.

Target	Gaussian	IMQ	polynomial	\mathbf{E} (exact)	RAY
coupled (needs both)	0.098	0.099	0.350	0.087	0.093
proximity only	0.341	0.141	0.612	0.237	0.278
alignment only	0.349	0.363	0.001	0.268	0.268

The preference survives fair per-kernel tuning and has the expected budget caveat: exact \mathbf{E} keeps the lead on the coupled target after hyperparameter tuning, but a small fixed feature budget can under-resolve the $O(d^2)$ modulation and let cheaper single-factor RFFs win. Those controls, together with the gating diagnostic that explains the effect, are in Appendix D.1. The main point is the scoped one Table 7 isolates: the kernel is

useful when the target actually couples alignment and proximity, not as a universal replacement for simpler kernels.

5.4 Large-scale primal training on HIGGS

RAY is a primal feature map, never forming a Gram. We close by exercising that property where it is forced: HIGGS (Baldi et al., 2014), 11,000,000 examples with $d = 28$ physics features, a scale at which the $N \times N$ Gram ($\sim 10^{14}$ entries) cannot be stored. Compressed Bernstein–Schur features train as a memory-flat streaming primal (peak 8.5 GB, no Gram) at this scale, with compression matching or beating exact modulation at every budget and decisively at the largest, where the exact polynomial starves the radial factor to a handful of draws whose seed spread is itself large; Gaussian RFF leads on AUC, as expected on a smooth detector task with no alignment \times proximity coupling to exploit. The point is streaming feasibility where the Gram is impossible, not peak AUC; the full sweep, protocol, and table are in Appendix D (Table 19).

6 Discussion

The biased \mathbf{E} -kernel fits neither random-feature template (it is neither shift-invariant nor a dot-product kernel (Proposition 3.1)) yet RAY linearizes it with operator-norm guarantees, and for the entire Bernstein–Schur *class* rather than a single kernel (Theorem 3.6). The Schur factorization is what makes this work: it turns an intractable product into two textbook objects, an exact finite modulation and one completely monotone radial factor, each randomized by its native tool. The matrix-Bernstein analysis (Theorem 4.2) then controls the error through the top eigenvalues and an intrinsic dimension, not the crude $N \max_{i,j}$ route, and the deployed doubly-randomized estimator inherits that guarantee up to a single tunable sketch term (Theorem 4.4). Whitening that argument at the ridge sharpens it into a deployment condition: the effective dimension $d_{\text{eff}}(\lambda) = \text{tr}(KA^{-1})$ is the *exact* intrinsic dimension of the whitened matrix variance, so $O((1 + \|P\|_{\text{op}}/\lambda) \log(d_{\text{eff}}/\delta))$ radial draws preserve the kernel-ridge coefficients and the objective value to constants (Theorem 4.8, Corollary 4.9), and tilting the radial draw by the whitened leverage $\bar{d}_\lambda(\theta) = \psi_\theta^\top (A^{-1} \circ P) \psi_\theta$, whose mean is exactly $d_{\text{eff}}(\lambda)$, improves the count to $O((1 + d_{\text{eff}}) \log(d_{\text{eff}}/\delta))$ (Theorem 4.10). This entire matrix-level guarantee set, operator norm through ridge condition, holds verbatim for every Bernstein–Schur kernel with the modulation Gram in place of the polynomial one (Theorem 4.13). The tensor-product closure these pieces rely on is classical (Aronszajn, 1950); the contribution is the factorization that brings it to bear on a kernel outside both templates, and the class-level concentration analysis of the result.

The modulation feature is interchangeable: Theorem 4.4 sees it only as PSD and an η -spectral approximation of P , so any unbiased or low-rank such feature (TensorSketch (Avron et al., 2014), random-Maclaurin products (Kar & Karnick, 2012), or anchor squares) plugs into the same bound. On the sphere, where the modulation is a dot-product kernel, anchor squares are exact with few anchors, the route of the spherical \mathbf{E} -attention of Luna et al. (2026); off the sphere they are biased ($\mathbb{E}_a(a^\top x)^2(a^\top w)^2 = \|x\|^2\|w\|^2 + 2(x^\top w)^2 \neq (x^\top w + b)^2$), so we default to an unbiased sketch. The choice is geometry, not analysis.

Computationally, via the Kronecker structure

$$z(x)^\top z(w) = \frac{(x^\top w + b)^2}{\varepsilon D} \sum_{j=1}^D \psi_{t_j}(x)^\top \psi_{t_j}(w), \quad (18)$$

a kernel evaluation costs $O(D(D' + d))$ without forming the M_b -dimensional feature. In primal form the estimator never materializes the $N \times N$ Gram, paying $O(N D d_b)$ feature memory; this beats the $O(N^2)$ Gram whenever $D d_b \ll N$. The polynomial dimension $d_b = O(d^2)$ is the only term scaling poorly in d . The deployed estimator sketches it to $m \ll d^2$ (Section 5.2), making the feature dimension Dm at the price of one tunable sketch term (Theorem 4.4).

Several caveats bound the method. The variance constant $(R^2 + b)^4$ exceeds the R^4 of Gaussian RFF (the price of the polynomial factor and the bias), and it is sharp, not a loose bound: it factors out of the variance of *any* exact-modulation estimator with equality (Proposition A.2), so the realized Gram error grows

weakly with d through that constant rather than the rate (Appendix D.1). The estimator needs bounded-norm inputs, since the unbounded numerator destabilizes the kernel on raw data, though the off-sphere ball (Section 5.1) shows exact spherical support is not required; the data-independent importance proposal is likewise suboptimal when distances spread widely. The $O(d^2)$ floor of the exact-modulation feature is removed by sketching (Theorem 4.4), worthwhile except at very low d or very high target accuracy, where one takes $m \rightarrow \infty$. And at moderate N and fixed representation size, adaptive Nyström can be more accurate, fitting landmarks to the observed sample; RAY does not aim to replace landmark methods, but to provide data-independent, unbiased, streaming features.

Several questions remain open. (1) *Risk*. The leverage count of Theorem 4.10 closes the draw-count side of the ridge question; a minimax excess-risk theorem now needs exactly two more pieces: a deployment-grade approximation of the tilt \bar{d}_λ with its error folded into the bound, and control of the approximate kernel’s effective dimension $d_{\text{eff}}(K_D^*)$ (Remark 4.11), the Rudi–Rosasco-style feature-covariance refinement. A matching lower bound, that plain $\text{Exp}(\varepsilon)$ sampling *requires* $\Omega(\|P\|_{\text{op}}/\lambda)$ draws on worst-case data (the empirical gap in Theorem 4.10’s validation points that way), would certify that the tilt is necessary and not merely sufficient. (2) *Is the prefactor fundamental?* Proposition A.2 pins the $(R^2 + b)^4$ variance prefactor with equality *within* the factored family (exact modulation times any unbiased radial estimate); whether any unbiased, data-independent feature map of $k_{\mathbf{E},b}$ can beat it, for instance by correlating the modulation and radial randomness so their errors cancel, is open. Sketching only adds variance (Proposition A.9) and normalizing changes the kernel (Proposition A.8), so a positive answer would have to leave the factored family without leaving unbiasedness; we conjecture the prefactor is optimal. (3) *The peaked regime*. Small ε with aligned data defeats both estimators we have: the trigonometric error grows with attention sharpness (Appendix D.1) and positive features have infinite variance precisely there (Proposition D.2). A finite-variance estimator for this regime, e.g. a FAVOR-type mechanism with data-dependent normalization, is open. (4) *Geometry*. Can the variance bound be tightened on low-dimensional manifolds, where the data occupies a thin shell and the intrinsic dimension of Theorem 4.2 is empirically tiny (Appendix D.1)?

7 Conclusion

We gave a random-feature scheme for *Bernstein–Schur kernels* (finite-feature modulations of completely monotone radial kernels) and its flagship, the biased \mathbf{E} -kernel $k_{\mathbf{E},b}(w, x) = (w^\top x + b)^2 / (\|w - x\|^2 + \varepsilon)$, a kernel that is neither shift-invariant nor a dot-product kernel and so fits neither standard random-feature template. The scheme reads the kernel as a Schur product and randomizes *both* factors: it samples the radial Bernstein–Widder scale $t \sim \text{Exp}(\varepsilon)$ and applies random Fourier features to the resulting Gaussian, and it sketches the modulation, so the deployed feature has dimension Dm , free of the $O(d^2)$ size of the exact polynomial feature. Keeping the modulation exact is the analyzable limit ($m \rightarrow \infty$): there the estimator is unbiased with $O((R^2 + b)^4 / (D\varepsilon^2))$ variance and a radial sample count free of explicit dimension for a fixed dataset, and a matrix-Bernstein operator-norm bound holds; conditioning on the sketch carries that guarantee to the doubly-randomized estimator up to a single tunable sketch term (Theorem 4.4). Whitening the matrix-Bernstein argument at the ridge upgrades the operator-norm bound to a high-probability kernel-ridge guarantee, in which the effective dimension $d_{\text{eff}}(\lambda)$ is the exact intrinsic dimension of the matrix variance and $O((1 + \|P\|_{\text{op}}/\lambda) \log(d_{\text{eff}}/\delta))$ radial draws preserve the ridge solution and objective value (Theorem 4.8); tilting the radial draw by a closed-form whitened leverage function, whose mean over the base law is exactly $d_{\text{eff}}(\lambda)$, improves the count to $O((1 + d_{\text{eff}}) \log(d_{\text{eff}}/\delta))$ (Theorem 4.10). Experiments confirm the $O(1/\sqrt{D})$ rate, the $(R^2 + b)^4$ bias law, and the matched-landmark off-sphere behavior, where $k_{\mathbf{E},b}$ is genuinely non-dot-product and no dot-product reduction is available. Cost-matched, adaptive data-dependent Nyström can be more accurate at moderate N , while RAY provides data-independent, unbiased, streaming primal features for a nonstationary product kernel. The same doubly-randomized construction applies unchanged to the whole Bernstein–Schur class, and so does the full matrix-level guarantee set, operator norm through ridge condition, with the modulation Gram in place of the polynomial one (Theorem 4.13); the modulation randomizer (sketch, random-Maclaurin, or anchor features) is a geometry-dependent choice, with anchor features the natural primitive on the sphere and an unbiased sketch off it.

Reproducibility. The code for every experiment, with the paper’s figures and a project page, is available online.¹ The per-table datasets, splits, seed counts, and swept grids are collected in Appendix E, and each result table states its own seed count and error bars in its caption.

References

- Thomas D. Ahle, Michael Kapralov, Jakob B. T. Knudsen, Rasmus Pagh, Ameya Velingker, David P. Woodruff, and Amir Zandieh. Oblivious sketching of high-degree polynomial kernels. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2020.
- Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.
- Haim Avron, Huy Nguyen, and David Woodruff. Subspace embeddings for the polynomial kernel. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
- Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Quasi-Monte Carlo feature maps for shift-invariant kernels. *Journal of Machine Learning Research*, 17(120):1–38, 2016.
- Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning (ICML)*, 2017.
- Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.
- Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5:4308, 2014.
- Gregory Beylkin and Lucas Monzón. Approximation by exponential sums revisited. *Applied and Computational Harmonic Analysis*, 28(2):131–149, 2010.
- Taha Bouhsine. Action at a distance: A universal reproducing kernel Hilbert space from polynomial alignment and IMQ distance. arXiv preprint, 2026.
- Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with Performers. In *International Conference on Learning Representations (ICLR)*, 2021.
- Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- Amit Daniely, Roy Frostig, Vineet Gupta, and Yoram Singer. Random features for compositional kernels. *arXiv preprint arXiv:1703.07872*, 2017.
- David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1166–1174, 2013.
- Purushottam Kar and Harish Karnick. Random feature maps for dot product kernels. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.
- Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, 2020.

¹<https://www.tahabouhsine.com/ray>

-
- Frances Y. Kuo and Dirk Nuyens. Application of quasi-Monte Carlo methods to elliptic PDEs with random diffusion coefficients. *Acta Numerica*, 25:225–356, 2016.
- Nicolas Langrené, Xavier Warin, and Pierre Gruet. A spectral mixture representation of isotropic kernels with application to random Fourier features. *arXiv preprint arXiv:2411.02770*, 2024.
- Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood: Approximating kernel expansions in loglinear time. In *International Conference on Machine Learning (ICML)*, 2013.
- Chun-Liang Li, Wei-Cheng Chang, Youssef Mroueh, Yiming Yang, and Barnabás Póczos. Implicit kernel learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.
- Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random Fourier features. *Journal of Machine Learning Research*, 22(108):1–51, 2021.
- José M. Luna, Taha Bouhsine, and Krzysztof Choromanski. SLAY: Geometry-aware spherical linearized attention with yat-kernel. *arXiv preprint arXiv:2602.04915*, 2026.
- Cameron Musco and Christopher Musco. Recursive sampling for the Nyström method. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.
- Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.
- Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.
- Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- Sami Remes, Markus Heinonen, and Samuel Kaski. Non-stationary spectral kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5):058301, 2012.
- Isaac J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39(4): 811–841, 1938.
- Issai Schur. Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 140:1–28, 1911.
- Zheyang Shen, Markus Heinonen, and Samuel Kaski. Harmonizable mixture kernels with variational Fourier features. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3273–3282, 2019.
- Jean-Francois Ton, Seth Flaxman, Dino Sejdinovic, and Samir Bhatt. Spatial mapping with Gaussian processes and nonstationary Fourier features. *Spatial Statistics*, 28:59–78, 2018.

-
- Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.
- Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.
- Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: A unified understanding for transformer’s attention via the lens of kernel. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- Jonas Wacker, Motonobu Kanagawa, and Maurizio Filippone. Improved random features for dot product kernels. *Journal of Machine Learning Research*, 25(235):1–75, 2024.
- David V. Widder. *The Laplace Transform*. Princeton University Press, 1941.
- Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, 2000.
- Andrew Gordon Wilson and Ryan P. Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning (ICML)*, 2013.
- Felix X. Yu, Ananda Theertha Suresh, Krzysztof M. Choromanski, Daniel N. Holtmann-Rice, and Sanjiv Kumar. Orthogonal random features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

A Proofs of the Approximation Guarantees

Both results rest on a single per-scale term $Y = (\psi_t(x)^\top \psi_t(w))(p(x)^\top p(w))$, whose polynomial factor is exact, $p(x)^\top p(w) = (x^\top w + b)^2/\varepsilon$, and whose Gaussian factor is an unbiased D' -sample estimate of $g_t(x, w)$. Two facts about it are used repeatedly: the a.s. bound $|\psi_t(x)^\top \psi_t(w)| \leq \|\psi_t(x)\| \|\psi_t(w)\| \leq 2$ (each $\|\psi_t\|^2 = \frac{2}{D'} \sum_\ell \cos^2 \leq 2$), so $|Y| \leq 2(R^2 + b)^2/\varepsilon$; and the second moment $\mathbb{E}[(\psi_t(x)^\top \psi_t(w))^2] \leq \mathbb{E}_T[e^{-2T\|x-w\|^2}] + O(1/D')$. Corollary A.6 needs no separate proof: it is $\|A\|_{\text{op}} \leq \|A\|_F \leq N \max_{ij} |A_{ij}|$ applied to Theorem A.4.

Theorem A.1 (Variance envelope). *Let $V_D = \text{Var}[z(x)^\top z(w)]$. Accounting for both the inner (D') and outer (D) sampling,*

$$V_D \leq \frac{(\|x\|^2 + b)^2(\|w\|^2 + b)^2}{D\varepsilon^2} \left(\frac{\varepsilon}{2\|x-w\|^2 + \varepsilon} + \frac{3}{2D'} \right) \leq \frac{(R^2 + b)^4}{D\varepsilon^2} \left(1 + \frac{3}{2D'} \right). \quad (19)$$

The $3/(2D')$ term is the inner random-Fourier-feature variance; it does not vanish at the recommended $D' = 1$ (where the constant is $5/2$), only as $D' \rightarrow \infty$.

Proof of Theorem A.1. Write $z(x)^\top z(w) = \frac{1}{D} \sum_j Y_j$; the Y_j are i.i.d., so $V_D = \frac{1}{D} \text{Var}[Y_1] \leq \frac{1}{D} \mathbb{E}[Y_1^2] = \frac{1}{D} (p(x)^\top p(w))^2 \mathbb{E}[(\psi_T(x)^\top \psi_T(w))^2]$, the polynomial factor being deterministic. It obeys $(p(x)^\top p(w))^2 = (x^\top w + b)^4/\varepsilon^2 \leq (\|x\|^2 + b)^2(\|w\|^2 + b)^2/\varepsilon^2$ by Cauchy–Schwarz and $b \geq 0$. For the Gaussian factor, the law of total variance over $\omega \mid T$ then T gives $\mathbb{E}[(\psi_T(x)^\top \psi_T(w))^2] = \mathbb{E}_T[g_T(x, w)^2] + \mathbb{E}_T[v_T]/D'$, where $v_t = \text{Var}_\omega[2 \cos(\omega^\top x + \beta) \cos(\omega^\top w + \beta)]$. The outer term is $\mathbb{E}_{T \sim \text{Exp}(\varepsilon)}[e^{-2T\|x-w\|^2}] = \varepsilon/(2\|x-w\|^2 + \varepsilon)$. For the inner term, $\mathbb{E}_{\omega, \beta}[(2 \cos(\omega^\top x + \beta) \cos(\omega^\top w + \beta))^2] = 1 + \frac{1}{2}e^{-4t\|x-w\|^2} \leq \frac{3}{2}$, so $v_t \leq \frac{3}{2}$ and $\mathbb{E}_T[v_T] \leq \frac{3}{2}$. Substituting and dividing by D gives (19); the second form uses $\varepsilon/(2\|x-w\|^2 + \varepsilon) \leq 1$. (The $1/\varepsilon^2$ comes from the polynomial normalization $p = \varepsilon^{-1/2}p_b$.) \square

Proposition A.2 (Sharpness of the variance prefactor). *For any unbiased estimator of $k_{\mathbf{E}, b}(x, w)$ of the factored form $\widehat{k} = (p(x)^\top p(w)) \widehat{h}$, with the modulation inner product kept exact (deterministic) and \widehat{h} any unbiased estimator of $h_\varepsilon(x, w)$,*

$$\text{Var}[\widehat{k}] = \frac{(x^\top w + b)^4}{\varepsilon^2} \text{Var}[\widehat{h}].$$

The prefactor $(x^\top w + b)^4/\varepsilon^2 \leq (R^2 + b)^4/\varepsilon^2$ is therefore common to the entire factored family and is attained with equality, not a slack upper bound: Theorem 4.1 is the instance with \hat{h} the flat radial estimator, and no sharper analysis can remove the prefactor while the modulation is kept exact. Lowering it requires leaving the family, by normalizing the modulation (Proposition A.8, a different, cosine-rescaled kernel) or randomizing it (Proposition A.9, which only adds variance).

Proof. The modulation factor $p(x)^\top p(w) = (x^\top w + b)^2/\varepsilon$ is deterministic, so it leaves the variance as its square; the inequality is Cauchy–Schwarz $(x^\top w + b)^2 \leq (\|x\|^2 + b)(\|w\|^2 + b) \leq (R^2 + b)^2$ with $b \geq 0$. \square

Proposition A.3 (Explicit D, D' variance and optimal allocation). *Let $v_t(x, w) = \text{Var}_\omega[2 \cos(\omega^\top x + \beta) \cos(\omega^\top w + \beta)]$ be the variance of one trigonometric feature pair at scale t , and set $V_{\text{out}} = \text{Var}_{T \sim \text{Exp}(\varepsilon)}[g_T(x, w)]$, $V_{\text{in}} = \mathbb{E}_{T \sim \text{Exp}(\varepsilon)}[v_T(x, w)]$. Then*

$$\text{Var}[z(x)^\top z(w)] = \frac{(x^\top w + b)^4}{\varepsilon^2} \left(\frac{V_{\text{out}}}{D} + \frac{V_{\text{in}}}{D D'} \right). \quad (20)$$

At a fixed feature budget $B = D D'$ the right side is minimized by $D' = 1$ (hence $D = B$) whenever $V_{\text{out}} > 0$: the inner term V_{in}/B is fixed by the budget, while the outer term V_{out}/D decreases as $1/D$. If $V_{\text{out}} = 0$ (for example at $x = w$), every allocation with the same $D D'$ ties.

Proof of Proposition A.3. The polynomial factor is deterministic, so $\text{Var}[Y_1] = (p(x)^\top p(w))^2 \text{Var}[\psi_T(x)^\top \psi_T(w)]$ with $(p(x)^\top p(w))^2 = (x^\top w + b)^4/\varepsilon^2$. By the law of total variance, conditioning on the scale T and then on the D' inner frequencies,

$$\text{Var}[\psi_T(x)^\top \psi_T(w)] = \mathbb{E}_T[\text{Var}_\omega(\psi_T(x)^\top \psi_T(w) \mid T)] + \text{Var}_T(\mathbb{E}_\omega[\psi_T(x)^\top \psi_T(w) \mid T]).$$

The conditional mean is $g_T(x, w)$, giving the outer term $\text{Var}_T(g_T) = V_{\text{out}}$. The conditional object is an average of D' i.i.d. feature pairs, so its variance is v_T/D' , giving the inner term $\mathbb{E}_T[v_T]/D' = V_{\text{in}}/D'$. Dividing by D (the D blocks are i.i.d.) yields (20). With $D D' = B$ fixed, $V_{\text{in}}/(D D') = V_{\text{in}}/B$ is constant and V_{out}/D is decreasing in D , so $D' = 1$ is optimal. \square

Theorem A.4 (Uniform Gram approximation). *For $X = \{x_1, \dots, x_N\}$ with $\|x_i\| \leq R$, the exact-modulation estimator with D outer samples satisfies, with probability at least $1 - \delta$,*

$$\sup_{i, j \in [N]} |z(x_i)^\top z(x_j) - k_{\mathbf{E}, b}(x_i, x_j)| \leq \frac{(R^2 + b)^2}{\varepsilon} \sqrt{\frac{8 \log(2N^2/\delta)}{D}}. \quad (21)$$

The constant uses the a.s. bound $|\psi_t(x)^\top \psi_t(w)| \leq 2$, not the expectation-level value 1.

Corollary A.5 (Sample complexity). $D = O((R^2 + b)^4 \varepsilon^{-2} \tau^{-2} \log(N/\eta))$ outer samples suffice for entrywise error at most τ with probability $1 - \eta$, independent of dimension d .

Corollary A.6 (Operator-norm control). *Since $\|A\|_{\text{op}} \leq \|A\|_F \leq N \max_{i, j} |A_{ij}|$, Theorem A.4 gives $\|K_D - K\|_{\text{op}} \leq N(R^2 + b)^2 \varepsilon^{-1} \sqrt{8 \log(2N^2/\delta)}/D$ with probability $1 - \delta$. The factor N is the price of this elementary route; Theorem 4.2 removes it via matrix concentration, replacing $N \max_{i, j} |\cdot|$ with the top eigenvalues of the kernel and modulation Gram matrices and an intrinsic dimension (which can still scale with N in the worst case).*

Proof of Theorem A.4. Fix (x_i, x_j) ; the estimator averages D i.i.d. terms $Y_k \in [-c, c]$ with $c = 2(R^2 + b)^2/\varepsilon$ by the a.s. bound above. Hoeffding’s inequality gives $\mathbb{P}(|z(x_i)^\top z(x_j) - k_{\mathbf{E}, b}| > s) \leq 2 \exp(-D s^2/(2c^2)) = 2 \exp(-D s^2 \varepsilon^2/(8(R^2 + b)^4))$. A union bound over the $\leq N^2$ pairs and solving $2N^2 \exp(\cdot) \leq \delta$ for s yields (21). \square

Lemma A.7 (Schur multiplier bound). *Let $A \succeq 0$ with $\max_i A_{ii} \leq a$. (a) If $P \succeq 0$, then $0 \preceq A \circ P \preceq a \|P\|_{\text{op}} I$. (b) If $E = E^\top$ is symmetric (not necessarily PSD), then $\|A \circ E\|_{\text{op}} \leq a \|E\|_{\text{op}}$.*

Proof. Write $A = UU^\top = \sum_r u_r u_r^\top$ with u_r the columns of U ; then $A \circ E = \sum_r (u_r u_r^\top) \circ E = \sum_r D_r E D_r$ with $D_r = \text{diag}(u_r)$, and the diagonal matrix $\sum_r D_r^2$ has entries $\sum_r (u_r)_i^2 = (UU^\top)_{ii} = A_{ii} \leq a$, so $\sum_r D_r^2 \preceq aI$. For any unit vector x and symmetric E ,

$$|x^\top (A \circ E)x| = \left| \sum_r (D_r x)^\top E (D_r x) \right| \leq \|E\|_{\text{op}} \sum_r \|D_r x\|^2 = \|E\|_{\text{op}} x^\top \left(\sum_r D_r^2 \right) x \leq a \|E\|_{\text{op}},$$

which is (b). Taking $E = P \succeq 0$ removes the absolute value (each summand is nonnegative), giving $0 \leq x^\top (A \circ P)x \leq a \|P\|_{\text{op}}$ and $A \circ P \succeq 0$ by the Schur product theorem, which is (a). \square

Proof of Theorem 4.2. Write $K_D - K = \sum_{j=1}^D Y_j$ with $Y_j = D^{-1}(K^{(j)} - K)$ and $K^{(j)} = (\Psi_j \Psi_j^\top) \circ P$ the Gram of the j -th radial draw; the Y_j are i.i.d. symmetric with $\mathbb{E}[Y_j] = 0$ (Theorem 3.4). Two structural facts: $K^{(j)} \succeq 0$, being a Schur product of the PSD matrices $\Psi_j \Psi_j^\top$ and P ; and $\|K^{(j)}\|_{\text{op}} \leq 2\|P\|_{\text{op}}$ by Lemma A.7 with $(\Psi_j \Psi_j^\top)_{ii} = \|\psi_{i_j}(x_i)\|^2 \leq 2$. Likewise $\|K\|_{\text{op}} = \|R \circ P\|_{\text{op}} \leq \|P\|_{\text{op}}$ by Lemma A.7, with R the radial Gram (PSD, unit diagonal).

A.s. bound. $\|Y_j\|_{\text{op}} \leq D^{-1}(\|K^{(j)}\|_{\text{op}} + \|K\|_{\text{op}}) \leq 3\|P\|_{\text{op}}/D =: L$.

Matrix variance. For PSD M , $M^2 \preceq \|M\|_{\text{op}} M$; with $\|K^{(j)}\|_{\text{op}} \leq 2\|P\|_{\text{op}}$ this gives $(K^{(j)})^2 \preceq 2\|P\|_{\text{op}} K^{(j)}$, so $\mathbb{E}[(K^{(j)})^2] \preceq 2\|P\|_{\text{op}} K$ and $\mathbb{E}[(K^{(j)} - K)^2] = \mathbb{E}[(K^{(j)})^2] - K^2 \preceq 2\|P\|_{\text{op}} K$. Hence

$$V = \sum_{j=1}^D \mathbb{E}[Y_j^2] = \frac{1}{D} \mathbb{E}[(K^{(1)} - K)^2] \preceq \frac{2\|P\|_{\text{op}}}{D} K, \quad \|V\|_{\text{op}} \leq \frac{2\|P\|_{\text{op}}\|K\|_{\text{op}}}{D} =: v.$$

Conclusion (expectation). We use the expectation form obtained by integrating the two-sided intrinsic tail below, rather than substituting d_{int} into the ambient-dimension expectation theorem. Let $\ell_0 = \log(8d_{\text{int}})$ and $s(u) = \sqrt{2vu} + \frac{2}{3}Lu$. The tail bound gives $\mathbb{P}\{\|\sum_j Y_j\|_{\text{op}} > s(\ell_0 + r)\} \leq e^{-r}$ for $r \geq 0$; integrating this quantile bound and using $\ell_0 \geq 1$ gives

$$\mathbb{E}\|\sum_j Y_j\|_{\text{op}} \leq s(\ell_0) + \int_0^\infty e^{-r} s'(\ell_0 + r) dr \leq 3\sqrt{v\ell_0} + 2L\ell_0.$$

Substituting $v = 2\|P\|_{\text{op}}\|K\|_{\text{op}}/D$ and $L = 3\|P\|_{\text{op}}/D$ yields the stated bound; $d_{\text{int}} \leq \text{rank}(V) \leq N$ and $\|P\|_{\text{op}} \leq \text{tr}(P) \leq N(R^2 + b)^2/\varepsilon$ give the worst-case comparison to Corollary A.6.

Conclusion (tail). With the *same* v and L , the tail form of the intrinsic matrix Bernstein inequality (Tropp, 2015, Thm. 7.3.1) bounds the top eigenvalue. Applying it to both $\sum_j Y_j$ and $-\sum_j Y_j$ and union bounding gives $\mathbb{P}\{\|\sum_j Y_j\|_{\text{op}} \geq s\} \leq 8d_{\text{int}} \exp(-\frac{s^2/2}{v+Ls/3})$ for $s \geq \sqrt{v} + L/3$. Setting the right side to δ and writing $\ell = \log(8d_{\text{int}}/\delta)$, the quadratic $s^2 - \frac{2}{3}L\ell s - 2v\ell = 0$ has positive root $s = \frac{1}{3}L\ell + \sqrt{(\frac{1}{3}L\ell)^2 + 2v\ell} \leq \sqrt{2v\ell} + \frac{2}{3}L\ell$. Because $\ell \geq \log 8 > 1$, this root obeys the range restriction: $\sqrt{2v\ell} \geq \sqrt{v}$ and $\frac{2}{3}L\ell \geq L/3$. Hence, with probability at least $1 - \delta$, $\|\sum_j Y_j\|_{\text{op}} \leq \sqrt{2v \log(8d_{\text{int}}/\delta)} + \frac{2}{3}L \log(8d_{\text{int}}/\delta)$, which is (10) after substituting v and L (the linear coefficient is $\frac{2}{3}L = 2\|P\|_{\text{op}}/D$). \square

Proof of Theorem 4.1. Write $Y = 2 \cos(\omega^\top x + \beta) \cos(\omega^\top w + \beta) = \cos(\omega^\top (x - w)) + \cos(\omega^\top (x + w) + 2\beta)$. The second term has zero mean over $\beta \sim \text{Unif}[0, 2\pi]$, so with $\omega \mid T \sim \mathcal{N}(0, 2TI_d)$ and the Gaussian characteristic function $\mathbb{E}_\omega[\cos(\omega^\top u)] = e^{-T\|u\|^2}$, $\mathbb{E}[Y \mid T] = e^{-Tr}$ and $\mathbb{E}[Y] = \mathbb{E}_{T \sim \text{Exp}(\varepsilon)}[e^{-Tr}] = \varepsilon/(\varepsilon + r)$. For the second moment, $\mathbb{E}_\omega[Y^2 \mid \omega] = \cos^2(\omega^\top (x - w)) + \frac{1}{2}$, and $\mathbb{E}_\omega[\cos^2(\omega^\top (x - w)) \mid T] = \frac{1}{2}(1 + e^{-4Tr})$ (using $\cos^2 \theta = \frac{1}{2}(1 + \cos 2\theta)$ and $\mathbb{E}[\cos(2\omega^\top (x - w)) \mid T] = e^{-4Tr}$). Hence $\mathbb{E}[Y^2 \mid T] = 1 + \frac{1}{2}e^{-4Tr}$ and $\mathbb{E}[Y^2] = 1 + \frac{1}{2}\varepsilon/(\varepsilon + 4r)$. Thus $\text{Var}(Y) = 1 + \frac{1}{2}\varepsilon/(\varepsilon + 4r) - (\varepsilon/(\varepsilon + r))^2$, and \widehat{k}_D averages D i.i.d. copies scaled by a/ε , giving $\text{Var}[\widehat{k}_D] = (a/\varepsilon)^2 \text{Var}(Y)/D$. \square

Proof of Theorem 4.7. Let $B = A^{-1/2}EA^{-1/2}$, so $\|B\|_{\text{op}} \leq \rho$ gives $-\rho I \preceq B \preceq \rho I$, and multiplying by $A^{1/2}$ on both sides gives $-\rho A \preceq E \preceq \rho A$; since $K_D + \lambda I = A + E$ this is the spectral sandwich. For the coefficients, $\tilde{\alpha} - \hat{\alpha} = (A + E)^{-1}y - A^{-1}y = -(A + E)^{-1}E\hat{\alpha}$, so $A^{1/2}(\tilde{\alpha} - \hat{\alpha}) = -(I + B)^{-1}B A^{1/2}\hat{\alpha}$. As $\|B\|_{\text{op}} \leq \rho < 1$, $\|(I + B)^{-1}\|_{\text{op}} \leq (1 - \rho)^{-1}$, hence $\|\tilde{\alpha} - \hat{\alpha}\|_A = \|A^{1/2}(\tilde{\alpha} - \hat{\alpha})\|_2 \leq \frac{\rho}{1 - \rho} \|A^{1/2}\hat{\alpha}\|_2 = \frac{\rho}{1 - \rho} \|\hat{\alpha}\|_A$. \square

Proof of Theorem 4.8. Write $A^{-1/2}(K_D - K)A^{-1/2} = \sum_{j=1}^D \tilde{Y}_j$ with $\tilde{Y}_j = D^{-1}A^{-1/2}(K^{(j)} - K)A^{-1/2}$, i.i.d. symmetric with mean zero. The three ingredients of the proof of Theorem 4.2 whiten as follows.

(a) *A.s. bound.* $K^{(j)} \succeq 0$ with $\|K^{(j)}\|_{\text{op}} \leq 2\|P\|_{\text{op}}$ gives $0 \preceq A^{-1/2}K^{(j)}A^{-1/2} \preceq 2\|P\|_{\text{op}}A^{-1} \preceq (2\|P\|_{\text{op}}/\lambda)I$, and $K \preceq A$ gives $0 \preceq A^{-1/2}KA^{-1/2} \preceq I$; hence $\|\tilde{Y}_j\|_{\text{op}} \leq L_\lambda := (1 + 2\|P\|_{\text{op}}/\lambda)/D$.

(b) *Variance majorant.* For PSD M , $MA^{-1}M = M^{1/2}(M^{1/2}A^{-1}M^{1/2})M^{1/2} \preceq \lambda^{-1}\|M\|_{\text{op}}M$; with $M = K^{(j)}$ and $\|K^{(j)}\|_{\text{op}} \leq 2\|P\|_{\text{op}}$, $\mathbb{E}[K^{(j)}A^{-1}K^{(j)}] \preceq (2\|P\|_{\text{op}}/\lambda)K$. Since $\mathbb{E}[(K^{(j)} - K)A^{-1}(K^{(j)} - K)] = \mathbb{E}[K^{(j)}A^{-1}K^{(j)}] - KA^{-1}K \preceq \mathbb{E}[K^{(j)}A^{-1}K^{(j)}]$,

$$\sum_j \mathbb{E}[\tilde{Y}_j^2] = \frac{1}{D}A^{-1/2}\mathbb{E}[(K^{(1)} - K)A^{-1}(K^{(1)} - K)]A^{-1/2} \preceq \tilde{V} := \frac{2\|P\|_{\text{op}}}{\lambda D}A^{-1/2}KA^{-1/2}.$$

(c) *Intrinsic dimension.* $\|\tilde{V}\|_{\text{op}} = \frac{2\|P\|_{\text{op}}}{\lambda D}\kappa_\lambda$ and $\text{tr}(\tilde{V}) = \frac{2\|P\|_{\text{op}}}{\lambda D}d_{\text{eff}}(\lambda)$, so $\text{intdim}(\tilde{V}) = d_{\text{eff}}(\lambda)/\kappa_\lambda = \tilde{d}_\lambda$.

The intrinsic-dimension matrix Bernstein inequality (Tropp, 2015, Thm. 7.3.1) accepts the semidefinite majorant \tilde{V} ; applied to $\pm \sum_j \tilde{Y}_j$ (the union over the two tails giving the prefactor 8) and inverted by the quadratic-root computation in the proof of Corollary 4.3, it yields, with probability at least $1 - \delta$ and $\ell = \log(8\tilde{d}_\lambda/\delta)$, $\rho_D \leq \sqrt{2\|\tilde{V}\|_{\text{op}}\ell} + \frac{2}{3}L_\lambda\ell$, which is (13). For the count: $2\sqrt{\kappa_\lambda\|P\|_{\text{op}}\ell}/(\lambda D) \leq \rho_0/2$ iff $D \geq 16\kappa_\lambda\|P\|_{\text{op}}\ell/(\lambda\rho_0^2)$, and $\frac{2}{3}(1 + 2\|P\|_{\text{op}}/\lambda)\ell/D \leq \rho_0/2$ iff $D \geq \frac{4}{3}(1 + 2\|P\|_{\text{op}}/\lambda)\ell/\rho_0$; both follow from $D \geq 16\rho_0^{-2}(1 + \|P\|_{\text{op}}/\lambda)\ell$ using $\kappa_\lambda \leq 1$, $1 + 2s \leq 2(1 + s)$, and $\rho_0 \leq 1$, giving (14). \square

Proof of Theorem 4.10. The leverage identity. With $D_\psi = \text{diag}(\psi_\theta)$, $K^{(\theta)} = D_\psi P D_\psi$, so $\text{tr}(A^{-1}K^{(\theta)}) = \text{tr}(D_\psi A^{-1} D_\psi P) = \psi_\theta^\top (A^{-1} \circ P) \psi_\theta$. The matrix $A^{-1} \circ P$ is a Schur product of PSD matrices, hence PSD, so $\tilde{d}_\lambda \geq 0$; and $\mathbb{E}_\pi[\psi_\theta \psi_\theta^\top] = R$ (the unit-diagonal radial Gram) gives $\mathbb{E}_\pi[\tilde{d}_\lambda] = \text{tr}(A^{-1}(R \circ P)) = \text{tr}(A^{-1}K) = d_{\text{eff}}(\lambda)$, so π_λ^* is a probability law. On $\{\tilde{d}_\lambda = 0\}$ the PSD matrix $A^{-1/2}K^{(\theta)}A^{-1/2}$ has zero trace, hence is zero, so $K^{(\theta)} = 0$ ($A^{-1/2}$ is invertible) and discarding these draws loses nothing. *Unbiasedness:* $\mathbb{E}_{\pi_\lambda^*}[\frac{d_{\text{eff}}}{\tilde{d}_\lambda}K^{(\theta)}] = \int K^{(\theta)} \mathbb{1}\{\tilde{d}_\lambda > 0\} d\pi = K$.

Whitened ingredients. Set $M_\theta = \frac{d_{\text{eff}}}{\tilde{d}_\lambda(\theta)}A^{-1/2}K^{(\theta)}A^{-1/2} \succeq 0$ and $\tilde{Y}_j = D^{-1}(M_{\theta_j} - A^{-1/2}KA^{-1/2})$, i.i.d. with mean zero under π_λ^* . (a) *A.s. bound.* On the support of π_λ^* , $\|M_\theta\|_{\text{op}} \leq \text{tr}(M_\theta) = \frac{d_{\text{eff}}}{\tilde{d}_\lambda} \text{tr}(A^{-1}K^{(\theta)}) = d_{\text{eff}}$ (a PSD matrix's norm is at most its trace, and the tilt normalizes the trace exactly); with $\|A^{-1/2}KA^{-1/2}\|_{\text{op}} \leq 1$, $\|\tilde{Y}_j\|_{\text{op}} \leq L^* := (1 + d_{\text{eff}})/D$. (b) *Variance majorant.* $M_\theta^2 \preceq \|M_\theta\|_{\text{op}}M_\theta \preceq d_{\text{eff}}M_\theta$ and $\mathbb{E}_{\pi_\lambda^*}[M_\theta] = A^{-1/2}KA^{-1/2}$, so

$$\sum_{j=1}^D \mathbb{E}[\tilde{Y}_j^2] \preceq \frac{1}{D}\mathbb{E}_{\pi_\lambda^*}[M_\theta^2] \preceq \tilde{V}^* := \frac{d_{\text{eff}}}{D}A^{-1/2}KA^{-1/2}.$$

(c) *Intrinsic dimension.* \tilde{V}^* is the same core matrix $A^{-1/2}KA^{-1/2}$ as in the proof of Theorem 4.8, rescaled, so $\text{intdim}(\tilde{V}^*) = \tilde{d}_\lambda$ exactly and $\|\tilde{V}^*\|_{\text{op}} = d_{\text{eff}}\kappa_\lambda/D$.

The intrinsic-dimension matrix Bernstein inequality applied to $\pm \sum_j \tilde{Y}_j$ and inverted as in Corollary 4.3 gives, with probability at least $1 - \delta$ and $\ell = \log(8\tilde{d}_\lambda/\delta)$, $\rho_D \leq \sqrt{2\|\tilde{V}^*\|_{\text{op}}\ell} + \frac{2}{3}L^*\ell$, which is (15). For the count: $\sqrt{2d_{\text{eff}}\kappa_\lambda\ell}/D \leq \rho_0/2$ iff $D \geq 8d_{\text{eff}}\kappa_\lambda\ell/\rho_0^2$, and $\frac{2}{3}(1 + d_{\text{eff}})\ell/D \leq \rho_0/2$ iff $D \geq \frac{4}{3}(1 + d_{\text{eff}})\ell/\rho_0$; both follow from $D \geq 8\rho_0^{-2}(1 + d_{\text{eff}})\ell$ using $\kappa_\lambda \leq 1$ and $\rho_0 \leq 1$, giving (16). \square

Proof of Corollary 4.9. The sandwich and the coefficient bound are Theorem 4.7 on the event $\rho_D \leq \rho_0$. For the objective value, the representer computation gives the optimal value $\lambda y^\top (K + \lambda I)^{-1}y$ for the exact problem (substitute $\hat{\alpha} = (K + \lambda I)^{-1}y$: $\|K\hat{\alpha} - y\|^2 + \lambda \hat{\alpha}^\top K \hat{\alpha} = \lambda^2 y^\top A^{-2}y + \lambda y^\top A^{-1}KA^{-1}y = \lambda y^\top A^{-1}y$) and likewise $\lambda y^\top (K_D + \lambda I)^{-1}y$ for the feature-space ridge problem with Gram K_D . Operator anti-monotonicity of the inverse on $(1 - \rho_0)A \preceq K_D + \lambda I \preceq (1 + \rho_0)A$ gives $(1 + \rho_0)^{-1}A^{-1} \preceq (K_D + \lambda I)^{-1} \preceq (1 - \rho_0)^{-1}A^{-1}$; evaluating the quadratic form at y finishes. \square

Proof of Corollary 4.12. Proposition 4.6 gives, on the sketch event, the deterministic bound $\rho_{\text{sk}} := \|A^{-1/2}(K_S - K)A^{-1/2}\|_{\text{op}} \leq \eta$ for $K_S = \hat{P}_m \circ R$. Conditioned on the sketch, $\hat{K}_{D,m}$ is an exact-modulation estimate of K_S with modulation Gram \hat{P}_m and $\|\hat{P}_m\|_{\text{op}} \leq (1 + \eta)\|P\|_{\text{op}}$. Applying Theorem 4.8 to the conditioned pair (\hat{P}_m, K_S) , whitened by $A_S = K_S + \lambda I$, with target $\rho_0/2$ and intrinsic dimension $\tilde{d}_{\lambda,S}$ gives

$$\|A_S^{-1/2}(\hat{K}_{D,m} - K_S)A_S^{-1/2}\|_{\text{op}} \leq \rho_0/2$$

under the count in the statement. The ridge transfer also gives $A_S \preceq (1 + \eta)A$, so the same radial error measured in the A -whitened norm is at most $(1 + \eta)\rho_0/2$. Therefore

$$\|A^{-1/2}(\hat{K}_{D,m} - K)A^{-1/2}\|_{\text{op}} \leq (1 + \eta)\rho_0/2 + \eta \leq \rho_0,$$

where the last inequality follows from $\eta \leq \rho_0/4$ and $\rho_0 \leq 1$. The final claim is Corollary 4.9 applied on the combined event. \square

Proposition A.8 (Normalized estimator). *Let $q_b(x) = p_b(x)/(\|x\|^2 + b)$, so $\|q_b(x)\| = 1$ (for $b > 0$ this is defined on all of \mathbb{R}^d ; for $b = 0$ we restrict to domains excluding the origin). Then $q_b(x)^\top q_b(w) = (x^\top w + b)^2 / ((\|x\|^2 + b)(\|w\|^2 + b))$, and the normalized kernel $k_{\mathbb{E},b}(x, w) = q_b(x)^\top q_b(w) (\|x - w\|^2 + \varepsilon)^{-1}$ is positive definite. Its flat estimator (replace p_b by q_b) is unbiased with, by Theorem 4.1 and $\|q_b\| = 1$, $\text{Var}[\hat{k}_D(x, w)] \leq \frac{1}{D\varepsilon^2} (1 + \frac{1}{2} \frac{\varepsilon}{\varepsilon + 4\|x-w\|^2}) \leq \frac{3}{2D\varepsilon^2}$, a bound free of R and b .*

Proof of Proposition A.8. $\|p_b(x)\|^2 = (\|x\|^2 + b)^2$ (Proposition 3.2), so $q_b = p_b/(\|x\|^2 + b)$ has unit norm and $q_b(x)^\top q_b(w) = p_b(x)^\top p_b(w) / ((\|x\|^2 + b)(\|w\|^2 + b)) = (x^\top w + b)^2 / ((\|x\|^2 + b)(\|w\|^2 + b))$. The normalized kernel is a Schur product of this PSD kernel and h_ε , hence PSD; the variance bound is Theorem 4.1 with $a = q_b(x)^\top q_b(w) \leq 1$. \square

Proposition A.9 (RAY error decomposition). *Sketch only the quadratic term: $\hat{p}_m(x, w) = \text{TS}_2(x)^\top \text{TS}_2(w) + 2bx^\top w + b^2$ with $\mathbb{E}[\text{TS}_2(x)^\top \text{TS}_2(w)] = (x^\top w)^2$, so $\mathbb{E}[\hat{p}_m] = p = (x^\top w + b)^2$ (the linear and constant terms are kept exact and add no sketch variance). Let \hat{p}_m be independent of the unbiased radial estimate \hat{h}_D of $h(x, w) = (\|x - w\|^2 + \varepsilon)^{-1}$. Then $\hat{k}_{D,m} = \hat{p}_m \hat{h}_D$ is unbiased for $k_{\mathbb{E},b}$ and*

$$\text{Var}[\hat{k}_{D,m}] = \underbrace{p^2 \text{Var}[\hat{h}_D]}_{\text{radial Monte Carlo}} + \underbrace{h^2 \text{Var}[\hat{p}_m]}_{\text{polynomial sketch}} + \underbrace{\text{Var}[\hat{p}_m] \text{Var}[\hat{h}_D]}_{\text{interaction}}.$$

The degree-2 TensorSketch bound $\text{Var}[\hat{p}_m] \leq C\|x\|^4\|w\|^4/m$ (the bias terms being exact) makes the sketch terms vanish as $m \rightarrow \infty$, recovering the exact-modulation estimator.

Proof of Proposition A.9. By independence $\mathbb{E}[\hat{p}_m \hat{h}_D] = ph = k$ and $\mathbb{E}[\hat{p}_m^2 \hat{h}_D^2] = \mathbb{E}[\hat{p}_m^2] \mathbb{E}[\hat{h}_D^2] = (p^2 + \text{Var} \hat{p}_m)(h^2 + \text{Var} \hat{h}_D)$; subtracting $p^2 h^2$ gives the three-term identity. \square

Proposition A.10 (Optimal sketch size). *At a fixed feature budget $M = D(m + d + 1)$, the deployed variance of Proposition A.9 is, to leading order in the two sampling counts,*

$$V(m) = \underbrace{\frac{A(m + d + 1)}{M}}_{\text{radial, } A=p^2 V_{\text{rad}}} + \underbrace{\frac{B}{m}}_{\text{sketch, } B=h^2 C\|x\|^4\|w\|^4} + O\left(\frac{1}{M}\right),$$

with A the radial Monte-Carlo constant ($V_{\text{rad}} = \text{Var}[\hat{h}_1]$) and B the degree-2 TensorSketch constant. It is minimized at the interior point

$$m^* = \sqrt{\frac{BM}{A}}, \quad V(m^*) = \frac{2\sqrt{AB}}{\sqrt{M}} + \frac{A(d + 1)}{M},$$

so the optimal sketch size grows as $m^* \propto \sqrt{M}$ with the budget and as $\sqrt{B/A}$ with the sketch-to-radial constant ratio; its d -dependence enters only through that ratio, hence is slow.

Proof. Substitute $D = M/(m + d + 1)$ into the radial term A/D of Proposition A.9 and keep the budget-independent sketch term B/m ; then $dV/dm = A/M - B/m^2 = 0$ gives $m^* = \sqrt{BM/A}$, and substitution gives $V(m^*)$. \square

Fitting A, B to the deployed-allocation data behind Table 5 by this two-term law confirms the rule: it explains 88–90% of the squared-error variance (R^2), the predicted $m^* = \sqrt{BM/A}$ lands at or within one grid step of the empirical error minimum, and the ratio $\sqrt{B/A}$ rises only from 0.41 to 0.58 as d goes from 16 to 256, the slow d -dependence the rule predicts (Appendix E, `sketch_size_rule`).

Proposition A.11 (The IMQ factor is a finite difference in b). *For every x, w , every $b \geq 0$, and every step $h > 0$,*

$$\frac{k_{\mathbf{E}, b+2h}(w, x) - 2k_{\mathbf{E}, b+h}(w, x) + k_{\mathbf{E}, b}(w, x)}{2h^2} = \frac{1}{\|w - x\|^2 + \varepsilon} = h_\varepsilon(w, x), \quad (22)$$

exactly, with no limit required and all three biases $b, b + h, b + 2h$ inside the kernel's domain $b \geq 0$.

Proof. The numerator $(w^\top x + b)^2$ is a quadratic polynomial in b , so any second difference equals the constant $2h^2$; the forward difference keeps all biases in $b \geq 0$ (the centered form would require $b \geq h$). Dividing by the b -independent denominator $\|w - x\|^2 + \varepsilon$ yields h_ε . \square

B Kronecker Implementation

For kernel evaluation only the inner product is needed, and it factors as (18), computable in $O(D(D' + d))$ per pair. For the full Gram matrix, Algorithm 1 assembles it via Hadamard products.

Algorithm 1 RAY Gram-matrix approximation

Require: Data $X = \{x_1, \dots, x_N\}$, parameters D, D', ε, b

- 1: Draw $t_1, \dots, t_D \sim \text{Exp}(\varepsilon)$
 - 2: Compute polynomial Gram $P_{ij} = (x_i^\top x_j + b)^2/\varepsilon$ $O(N^2d)$
 - 3: $K_{\text{approx}} \leftarrow \mathbf{0}_{N \times N}$
 - 4: **for** $j = 1, \dots, D$ **do**
 - 5: Draw $\omega_{j,\ell} \sim \mathcal{N}(0, 2t_j I_d)$, $\beta_{j,\ell} \sim \text{Unif}([0, 2\pi])$
 - 6: Build $\Psi_j \in \mathbb{R}^{N \times D'}$, $\Psi_j[i, \ell] = \sqrt{2/D'} \cos(\omega_{j,\ell}^\top x_i + \beta_{j,\ell})$
 - 7: $K_{\text{approx}} += (\Psi_j \Psi_j^\top) \circ P/D$ $O(N^2D' + ND'd)$
 - 8: **end for**
 - 9: **return** K_{approx}
-

C Spectral Interpretation: Bernstein–Schur Kernels

The scheme draws t from the Bernstein–Widder mixing measure of the completely monotone radial factor h_ε . This is not a Bochner spectral measure ($k_{\mathbf{E}, b}$ is not shift-invariant) but plays the analogous role, and the construction depends on h_ε only through this measure. Here we prove the general Theorem 3.6.

Proof of Theorem 3.6. For one draw set $Y = 2 \cos(\omega^\top x + \beta) \cos(\omega^\top w + \beta)$; conditioned on T the standard RFF identity gives $\mathbb{E}[Y | T] = e^{-T\|x-w\|^2}$, so $\mathbb{E}[m_f Y] = m_f \mathbb{E}_{T \sim \nu/m_f} [e^{-T\|x-w\|^2}] = \int_0^\infty e^{-t\|x-w\|^2} d\nu(t) = f(\|x-w\|^2)$. With $u(x)^\top u(w) = p(x, w)$, $\mathbb{E}[z(x)^\top z(w)] = p(x, w)f(\|x-w\|^2) = k(x, w)$. For the variance, $|m_f Y u(x)^\top u(w)| \leq 2m_f B^2$ and $\mathbb{E}[Y^2 | T] \leq \frac{3}{2}$ (as in the proof of Theorem 4.1), so $\text{Var}[m_f Y u(x)^\top u(w)] \leq m_f^2 B^4 \mathbb{E}[Y^2] \leq \frac{3}{2} m_f^2 B^4$; averaging D i.i.d. copies divides by D . For the uniform bound, each summand lies in an interval of length $4m_f B^2$, so Hoeffding gives $\mathbb{P}(|z(x_i)^\top z(x_j) - k(x_i, x_j)| \geq s) \leq 2 \exp(-Ds^2/(8m_f^2 B^4))$, and a union bound over $\leq N^2$ pairs yields the claim. \square

The biased \mathbf{E} -kernel is the case $u = p_b$ ($d_p = d(d+1)/2 + d + 1$), $f(r) = (r + \varepsilon)^{-1}$ with $d\nu(t) = e^{-\varepsilon t} dt$, so $m_f = 1/\varepsilon$ and $\nu/m_f = \text{Exp}(\varepsilon)$; nontrivial members are tabulated in the main text (Table 1).

Proof of Theorem 4.13. The class estimator of Theorem 3.6 has per-draw Gram $K^{(j)} = (\psi_j \psi_j^\top) \circ P_u$ with $\psi_j \in \mathbb{R}^N$, $\psi_j[i] = \sqrt{2} \cos(\omega_j^\top x_i + \beta_j)$, and $P_u = m_f G_u$ (the mass m_f entering through the $\sqrt{m_f/D}$ scaling of the feature). The four cited proofs consume exactly three structural facts about the triple $(K^{(j)}, P, K)$; we verify each with P_u in place of P .

(i) *PSD per-draw Gram with bounded operator norm.* $\psi_j \psi_j^\top \succeq 0$ is rank one with diagonal $2 \cos^2(\omega_j^\top x_i + \beta_j) \leq 2$, and $G_u = [u(x_i)^\top u(x_j)] \succeq 0$ is a Gram matrix, so $K^{(j)} \succeq 0$ by the Schur product theorem and $\|K^{(j)}\|_{\text{op}} \leq 2\|P_u\|_{\text{op}}$ by Lemma A.7(a) with $a = 2$.

(ii) *Unbiasedness through a unit-diagonal PSD radial Gram.* Conditional on T_j , the RFF identity gives $\mathbb{E}[(\psi_j \psi_j^\top)_{ik} \mid T_j] = e^{-T_j \|x_i - x_k\|^2}$; averaging over $T_j \sim \nu/m_f$ yields $\mathbb{E}[K^{(j)}] = R_u \circ P_u = K$ with $(R_u)_{ik} = f(\|x_i - x_k\|^2)/m_f$. The radial Gram R_u is PSD (a nonnegative mixture of Gaussian Grams under ν/m_f) with unit diagonal ($f(0) = m_f$), so Lemma A.7(a) also gives $K \succeq 0$ and $\|K\|_{\text{op}} \leq \|P_u\|_{\text{op}}$.

(iii) *The variance step.* For PSD M , $M^2 \preceq \|M\|_{\text{op}} M$ and $MA^{-1}M \preceq \lambda^{-1} \|M\|_{\text{op}} M$; with (i) these give $\mathbb{E}[(K^{(j)} - K)^2] \preceq 2\|P_u\|_{\text{op}} K$ and $\mathbb{E}[(K^{(j)} - K)A^{-1}(K^{(j)} - K)] \preceq (2\|P_u\|_{\text{op}}/\lambda)K$, the two variance majorants.

Substituting (i)–(iii) into the proofs of Theorem 4.2 and Corollary 4.3 (a.s. bound $L = 3\|P_u\|_{\text{op}}/D$, variance $v = 2\|P_u\|_{\text{op}}\|K\|_{\text{op}}/D$) and of Theorem 4.8 and Corollary 4.9 (whitened a.s. bound $L_\lambda = (1 + 2\|P_u\|_{\text{op}}/\lambda)/D$, majorant $\frac{2\|P_u\|_{\text{op}}}{\lambda D} A^{-1/2} K A^{-1/2}$, whose intrinsic dimension is again exactly \tilde{d}_λ) reproduces every statement verbatim with $P \mapsto P_u$. The leverage theorem transfers identically: its proof consumes, beyond (i)–(iii), only the identity $\mathbb{E}_\pi[\psi \psi^\top] = R_u$ (the unit-diagonal radial Gram of (ii)), so $\tilde{d}_\lambda(\theta) = \psi_\theta^\top (A^{-1} \circ P_u) \psi_\theta$ has mean $d_{\text{eff}}(\lambda)$ and Theorem 4.10 holds with $P \mapsto P_u$. For $k_{\mathbf{E},b}$, $u = p_b$ and $m_f = 1/\varepsilon$ give $P_u = [(x_i^\top x_j + b)^2/\varepsilon] = P$. \square

For the polynomially modulated Matérn- $\frac{1}{2}$ kernel $(x^\top w + b)^q e^{-\|x-w\|/\sigma}$ of Theorem 4.13: the radial factor $f(r) = e^{-\sqrt{r}/\sigma}$ is completely monotone in $r = \|x - w\|^2$ with the Lévy/inverse-Gaussian Bernstein representation

$$e^{-\sqrt{r}/\sigma} = \int_0^\infty \frac{1}{2\sigma\sqrt{\pi}} t^{-3/2} e^{-1/(4\sigma^2 t)} e^{-tr} dt,$$

mass $m_f = f(0) = 1$, and exact sampler $T = 1/(2\sigma^2 Z^2)$, $Z \sim \mathcal{N}(0, 1)$ (the Lévy law of parameter $c = 1/(2\sigma^2)$ is the law of c/Z^2); the modulation Gram is $G_u = [(x_i^\top x_j + b)^q]$ with $\|u(x)\|^2 = (\|x\|^2 + b)^q \leq (R^2 + b)^q$, so $P_u = G_u$ and every bound of Theorem 4.13 applies with $\|P_u\|_{\text{op}} \leq N(R^2 + b)^q$ in the worst case and the data-adaptive $\|P_u\|_{\text{op}}$ in general.

D Further Validations

D.1 Experiment controls and scoped deployment checks

Fair-cost comparison. Matched draws understate RAY’s cost; matched representation is the honest axis. The Nyström methods here are strong *data-dependent* deployment baselines (Table 3), not data-independent feature maps. On the off-sphere bounded ball (Table 8; $d = 64$, $\|x\| \in [0.3, 1.5]$, coupled target, $M = d_b = 2145$), RAY far outperforms the budget-starved exact-modulation limit, but adaptive k-means and ridge-leverage Nyström of the *exact* \mathbf{E} -kernel are most accurate. This is the main deployment caveat: at moderate N and fixed representation size, adaptive landmarks can be better unless streaming, unbiasedness, or pre-data features are required.

Tuned and matched-budget controls for the coupling test. The coupled-target preference of Table 7 survives fair per-kernel tuning (Table 9): with every kernel’s b, ε, λ grid-searched on a held-out split, the exact \mathbf{E} -kernel is best on the coupled target and statistically tied on **digits**, where one factor already suffices.

At small fixed feature dimension the story changes because resolving the coupling through random features costs the polynomial floor. Table 10 shows this explicitly: the exact \mathbf{E} kernel keeps the lead on the coupled target, but sketched RAY is under-resolved at $M = 4096$ and a cheaper single-factor RFF can score better.

Table 8: Off-sphere fair-cost: KRR test RMSE (\downarrow) on a bounded ball ($d = 64$, $\|x\| \in [0.3, 1.5]$, coupled target, mean \pm std over 3 seeds) at matched representation dimension. RAY beats the budget-starved exact-modulation limit; k-means and ridge-leverage Nyström of the exact \mathbf{E} -kernel are most accurate; Gaussian RFF approximates a different kernel. The sphere-normalized version is Table 23.

Method	dim	RMSE	memory (MB)	build (s)
exact modulation	2145	1.114 \pm .360	25.7	0.020
RAY ($m=128$)	2123	0.473 \pm .010	25.5	0.015
k-means Nyström	1500	0.099 \pm .003	18.0	1.887
rls-leverage Nyström	1500	0.099 \pm .003	18.0	0.928
Gaussian RFF (different kernel)	2145	0.260 \pm .003	25.7	0.024

Table 9: Validation-tuned downstream comparison (every kernel’s b, ε, λ grid-searched on a held-out split; off-sphere; mean \pm std over 3 seeds). The \mathbf{E} -kernel’s coupled-target advantage survives fair tuning: it wins the synthetic coupled (alignment \times proximity) target and ties on `digits` where one factor suffices. RAY is the deployed sketched approximation at a fixed budget.

kernel	coupled (RMSE \downarrow)	digits (acc \uparrow)
Gaussian	0.038 \pm .001	0.980 \pm .007
IMQ	0.036 \pm .002	0.979 \pm .004
polynomial	0.114 \pm .008	0.982 \pm .006
\mathbf{E} (exact)	0.034 \pm .001	0.978 \pm .004
RAY (sketched)	0.078 \pm .011	0.974 \pm .005
Nyström- \mathbf{E}	0.043 \pm .003	0.979 \pm .003

This does not negate the kernel preference; it identifies the regime where the feature construction is meant to be used.

Table 10: Matched-dimension random-feature comparison (RMSE \downarrow , off-sphere $d=32$, $M=4096$, mean of 3 seeds). At this small fixed budget the $O(d^2)$ polynomial floor leaves the coupling under-resolved, so a cheaper RFF for a simpler kernel scores better while the exact \mathbf{E} -kernel keeps the lead; RAY’s regime is large- N streaming, not a fixed-budget bake-off.

target	RAY (sketched)	Gaussian RFF	IMQ RFF	poly sketch	exact \mathbf{E}
coupled	0.103	0.038	0.039	0.143	0.025
proximity	0.477	0.214	0.102	0.447	0.061
alignment	0.131	0.102	0.108	0.016	0.088

The alignment numerator as signal modulation. The coupled-target preference is explained by the modulation–radial decomposition (Proposition 4.14): radial Monte-Carlo noise enters each Gram entry scaled by the alignment modulation p_b , so the product sharpens *between*-pair discrimination while leaving any single pair’s relative error unchanged. On off-sphere data ($d = 16$) we score three pair types (*true* (close and aligned), *radial distractors* (close but weakly aligned), and *alignment distractors* (aligned but far)) by the AUC separating true from each (400 pairs/type). A radial-only kernel (IMQ) false-positives on radial distractors (AUC 0.22); an alignment-only kernel (degree-2 polynomial) false-positives on alignment distractors (AUC 0.00); only the \mathbf{E} product suppresses both (AUC 1.00/1.00), the exact-modulation estimator inherits this almost exactly, and the deployed RAY (sketched modulation) partially survives on the alignment distractors (AUC 0.73).

Linear-time streaming attention. Attention is a kernel smoother over tokens, $\text{attn}(q_i) = \sum_j k(q_i, k_j)v_j / \sum_j k(q_i, k_j)$, at $O(N^2)$ cost. An explicit feature map collapses it to $O(NM)$; with ϕ such

that $\phi(q)^\top \phi(k) \approx k(q, k)$, the smoother factorizes as $\phi(q_i)^\top (\sum_j \phi(k_j) v_j^\top) / \phi(q_i)^\top \sum_j \phi(k_j)$. RAY supplies such a map for $k_{\mathbb{E}, b}$ and is causal through the recurrent state $S_t = S_{t-1} + \phi(k_t) v_t^\top$. A landmark method cannot serve here because its landmarks are chosen from the sequence.

We measure fidelity and scaling, not language quality. The linear-attention output matches exact \mathbb{E} -attention with a median per-token error that falls with the feature dimension M (Figure 3a, $d = 32$: 0.45 at $M=1552$ down to 0.12 at $M=12416$), and the induced attention-weight matrix tracks it ($0.57 \rightarrow 0.12$). The approximation does not drift upward with context length, but it does depend on sharpness: diffuse attention is easy, peaked attention is the hard regime for the radial RFF (Figure 3b). The payoff is the $O(N^2)$ wall (Figure 3c): the exact $N \times N$ attention matrix reaches 137 GB at $N=131072$ while RAY stays linear and a single recurrent state decodes causally.

The sign indefiniteness has a principled fix, and whether that fix also cures the peaked regime has a sharp, negative answer. The trigonometric radial feature is the source both of the peaked-regime error and of an estimator that can return negative attention weights; a positive feature removes the second problem (Proposition D.1), and Proposition D.2 shows it cannot remove the first.

Proposition D.1 (Positive features give nonnegative attention). *The Gaussian radial factor admits a positive random feature: for $t > 0$ and $\omega \sim \mathcal{N}(0, 2tI_d)$,*

$$e^{-t\|x-w\|^2} = \mathbb{E}_\omega[\phi_t^+(x) \phi_t^+(w)], \quad \phi_t^+(x) = \exp(\omega^\top x - 2t\|x\|^2) > 0, \quad (23)$$

the FAVOR⁺ construction of Choromanski et al. (2021), with controlled relative error where the kernel is small. Combined with a nonnegative modulation feature, anchor squares $\phi_a(x) = (a^\top x)^2 \geq 0$ (exact on the sphere, Section 6), the product feature $\Phi(x) = \phi_t^+(x) \phi_a(x)$ is entrywise positive, so the estimated \mathbb{E} -attention weights $\Phi(q)^\top \Phi(k) \geq 0$ are nonnegative by construction, which the signed trigonometric/TensorSketch estimator cannot guarantee.

Proof. For $\omega \sim \mathcal{N}(0, 2tI_d)$, $\mathbb{E}_\omega[e^{\omega^\top(x+w)}] = e^{t\|x+w\|^2}$, so $\mathbb{E}_\omega[\phi_t^+(x) \phi_t^+(w)] = e^{-2t\|x\|^2 - 2t\|w\|^2} e^{t\|x+w\|^2} = e^{-t\|x-w\|^2}$, which is (23); positivity of ϕ_t^+ and of ϕ_a is immediate, and a product of nonnegative features has a nonnegative inner product. \square

Proposition D.2 (Variance dichotomy: the mixing law caps positive features). *Let $\hat{g}^+ = \phi_T^+(x) \phi_T^+(w)$ with $T \sim \text{Exp}(\varepsilon)$, $\omega \mid T \sim \mathcal{N}(0, 2TI_d)$, the positive-feature estimate of the rescaled radial factor. Then $\mathbb{E}[\hat{g}^+] = \varepsilon / (\|x-w\|^2 + \varepsilon)$, but*

$$\mathbb{E}[(\hat{g}^+)^2] = \mathbb{E}_T[e^{8Tx^\top w}] = \begin{cases} \frac{\varepsilon}{\varepsilon - 8x^\top w}, & 8x^\top w < \varepsilon, \\ +\infty, & 8x^\top w \geq \varepsilon, \end{cases}$$

so the positive-feature radial estimator has infinite variance on every pair with $x^\top w \geq \varepsilon/8$. Truncating the scale at $T \leq T_{\max}$ (relative bias at most $e^{-\varepsilon T_{\max}}$ on the radial factor) caps the second moment at $e^{8T_{\max} x^\top w}$, which remains exponential in $x^\top w/\varepsilon$. The trigonometric estimator instead satisfies $\mathbb{E}[\hat{g}^2 \mid T] = 1 + \frac{1}{2} e^{-4T\|x-w\|^2} \leq \frac{3}{2}$ uniformly in the scale and the pair.

Proof. Conditionally on $T = t$, $(\hat{g}^+)^2 = e^{2\omega^\top(x+w)} e^{-4t(\|x\|^2 + \|w\|^2)}$, and $\mathbb{E}_\omega[e^{a^\top \omega}] = e^{t\|a\|^2}$ for $\omega \sim \mathcal{N}(0, 2tI_d)$ gives $\mathbb{E}_\omega[(\hat{g}^+)^2 \mid t] = e^{4t\|x+w\|^2 - 4t\|x\|^2 - 4t\|w\|^2} = e^{8tx^\top w}$. Integrating against $\varepsilon e^{-\varepsilon t} dt$ gives $\varepsilon / (\varepsilon - 8x^\top w)$ when $8x^\top w < \varepsilon$ and diverges otherwise; on $\{T \leq T_{\max}\}$ the conditional bound $e^{8tx^\top w} \leq e^{8T_{\max} x^\top w}$ (for $x^\top w \geq 0$; for $x^\top w < 0$ it is at most 1) caps the second moment, and the truncated integral $\int_0^{T_{\max}} \varepsilon e^{-\varepsilon t} e^{-tr} dt$ misses at most $e^{-(\varepsilon+r)T_{\max}} \leq e^{-\varepsilon T_{\max}}$ of the total in relative terms. The trigonometric bound is in the proof of Theorem 4.1. \square

Empirically, the dichotomy is sharp: below the threshold the finite-side prediction is approached, while above it the empirical second moment grows without plateau and the positive-feature Gram estimator becomes heavy-tailed or catastrophic; the signed trigonometric estimator keeps the Monte-Carlo rate. Positive features are therefore appropriate only in diffuse regimes where the radial RFF is already accurate and nonnegative weights matter.

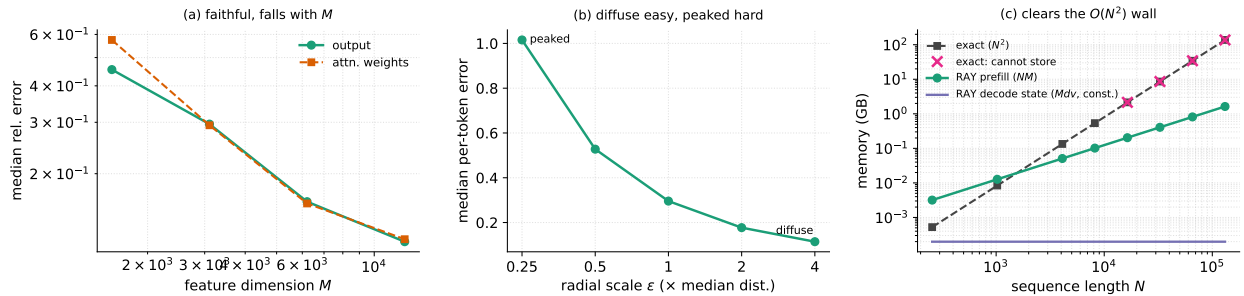


Figure 3: RAY as a linear-time, streaming \mathbf{E} -attention primitive (random queries/keys/values, $d = 32$). (a) The linear-attention output and the induced attention-weight matrix both match exact \mathbf{E} -attention with a median error that falls with the feature dimension M ; one fixed map is applied to every token. (b) The one limitation: error scales with attention sharpness: diffuse attention (large radial scale ϵ) is easy, peaked attention is the hard regime for the radial RFF. (c) The exact $N \times N$ attention matrix becomes impossible to store (137 GB at $N=131072$, \times), while RAY’s $O(NM)$ features stay linear in N and a single $O(Md_v)$ recurrent state (constant in N) decodes causally; that recurrence $S_t = S_{t-1} + \phi(k_t)v_t^\top$ is exact, which a data-dependent landmark scheme cannot offer.

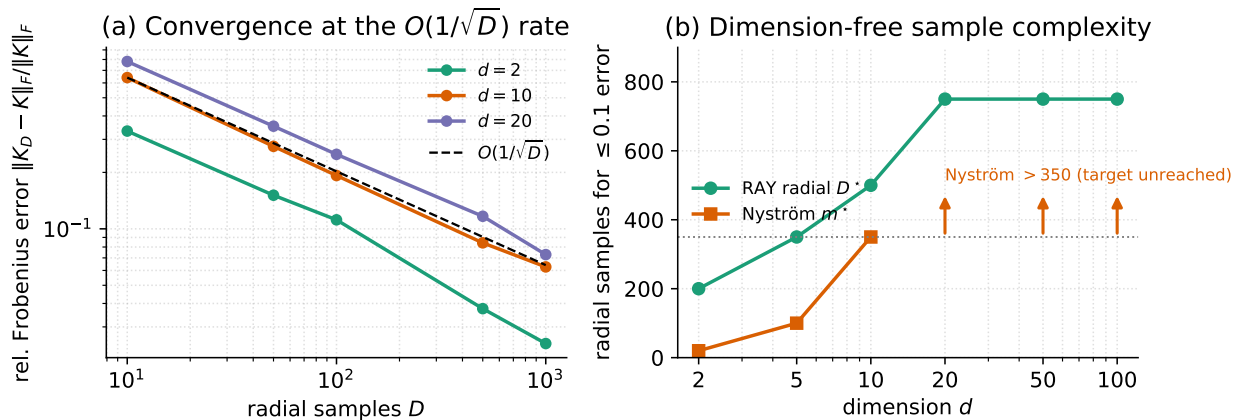


Figure 4: Sphere-normalized sanity check (here the kernel coincides with a dot-product kernel, so this isolates the dimension behavior and is not a representation claim). RAY approximates the biased Gram at the Monte-Carlo rate with a radial sample count that grows little with dimension (flat $D^* = 1$). (a) Relative Frobenius error vs. D ($N = 1000$, $b = 1$, $\epsilon = 1$); all dimensions track the $O(1/\sqrt{D})$ guide within a factor of ~ 3 . (b) Radial samples for $\leq 10\%$ error vs. dimension: RAY’s D^* stays bounded and plateaus by $d = 20$, while the Nyström landmark count m^* exceeds the tested range for $d \geq 20$.

Gram-matrix approximation error. Random \mathbf{E} -features converge at the Monte-Carlo rate, and their accuracy grows only mildly with dimension. Approximating the exact biased Gram matrix of $N = 1000$ sphere-normalized points (normalization fixes $R = 1$, isolating the dimension behavior from the R^2 growth of raw data; $b = 1$, $\epsilon = 1$, mean over 5 seeds, the recommended flat $D^* = 1$ so D is the number of cosine draws), the relative Frobenius error falls as $O(1/\sqrt{D})$ in the number of radial samples D (from 0.33 at $D = 10$ to 0.024 at $D = 1000$ at $d = 2$ (Figure 4a)) and at $D = 1000$ stays within $[0.024, 0.073]$ across $d \in \{2, \dots, 20\}$, a factor of ~ 3 . Uniform-landmark Nyström at a fixed budget behaves oppositely (Table 11; we add stronger k-means and adaptive ridge-leverage-score Nyström (Musco & Musco, 2017) baselines in Section 5.1): at $d = 2$ the error is below the table’s displayed precision, likely reflecting unusually rapid spectral decay on this sampled problem rather than literal exact recovery, but by $d = 20$ it degrades to 0.39 ($m=50$) and 0.29 ($m=100$). At matched draws versus landmarks, RAY is therefore preferable in moderate-to-high dimension, Nyström only in low. The operator-norm error controlled by Theorem 4.2 tracks the Frobenius error: the

same $O(1/\sqrt{D})$ rate, and smaller in relative terms ($0.43 \rightarrow 0.059$ vs. $0.59 \rightarrow 0.071$ as D runs $10 \rightarrow 1000$ at $d = 10$). The theorem’s data-adaptive constant is borne out directly: across bounded-ball datasets of increasing spectral spread (intrinsic dimension d_{int} from 1.4 to 3.5, $N = 400$) the matrix-Bernstein expression stays near 55–60 while the crude $N \max_{ij}$ route grows from 66 to 147, the gap widening with the spectrum, and both upper-bound the measured $\|K_D - K\|_{\text{op}}$ (20–24).

Table 11: Relative Frobenius error $\|K_{\text{approx}} - K\|_F / \|K\|_F$ of the biased Gram matrix ($b = 1$, $\varepsilon = 1$, unit sphere, mean over 5 seeds). RAY uses D radial samples at the recommended flat $D' = 1$, so D is the number of cosine draws (total feature dimension $D d_b$); Nyström uses m landmarks.

Method	$d = 2$	$d = 5$	$d = 10$	$d = 20$
RAY, $D=100$	0.112	0.174	0.192	0.250
RAY, $D=500$	0.038	0.076	0.084	0.117
RAY, $D=1000$	0.024	0.045	0.063	0.073
Nyström, $m=50$	0.000	0.188	0.332	0.393
Nyström, $m=100$	0.000	0.081	0.216	0.289

Bias scaling. The bias enters the variance exactly as the fourth power Theorem A.1 predicts (Figure 5). For two unit-vector pairs in \mathbb{R}^2 (aligned ($x^\top w = 1$) and $x^\top w = 0.5$) we sweep $b \in \{0, \dots, 5\}$ at $D = 200$, $\varepsilon = 1$ and fit the log-log slope of the empirical variance (2000 repetitions) against $x^\top w + b$.

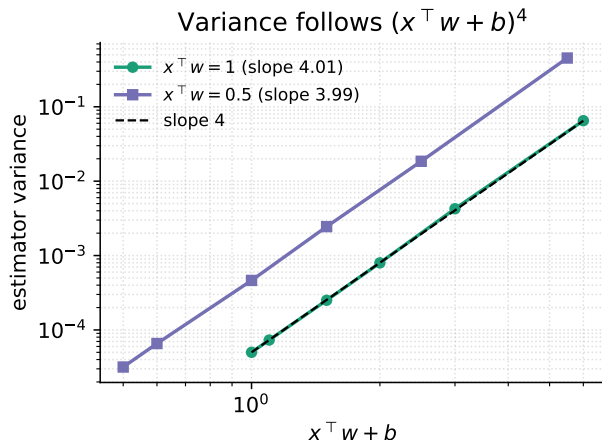


Figure 5: Estimator variance vs. the bias-shifted alignment $x^\top w + b$ (log-log, 2000 repetitions). Both pairs follow a fourth power (fitted slopes 4.01 and 3.99 against the slope-4 guide). For the aligned pair the variance equals the $(R^2 + b)^4$ envelope of Theorem A.1 (the ratio $\text{Var} / (R^2 + b)^4$ is constant at $\approx 5 \times 10^{-5}$, so the bound is tight); the $x^\top w = 0.5$ pair lies below it, the gap being the Cauchy–Schwarz step in the proof.

The fitted exponents are 4.01 and 3.99 (and 2.01, 1.99 for the standard deviation), matching the predicted 4 and 2. The aligned pair sits on the $(R^2 + b)^4$ envelope, where $\text{Var} / (R^2 + b)^4$ is constant and the bound is tight; the $x^\top w = 0.5$ pair lies strictly below it, the gap being the Cauchy–Schwarz step $x^\top w \leq \|x\| \|w\|$ in the proof.

Beyond the envelope, the *exact* variance of Theorem 4.1 is confirmed pointwise: across $b \in \{0, 1, 2\}$ and 18 off-sphere pairs spanning $r = \|x - w\|^2 \in [0.4, 2.2]$, the ratio of empirical (over 2×10^5 draws) to predicted variance is 1.001 ± 0.002 , and the estimator is unbiased to within 0.5%.

The normalized estimator removes the bias/radius variance blow-up. The normalized variant (Proposition A.8) trades $k_{\mathbb{E},b}$ for a bounded-variance rescaling. We confirm the exact relation $\text{Var}(\text{exact modulation}) / \text{Var}(\text{normalized}) = ((\|x\|^2 + b)(\|w\|^2 + b))^2$ by sweeping b at $R = 1$ and the data

radius R at $b = 1$ (2×10^5 draws per point, Table 12). The measured ratio matches $(R^2 + b)^4$ to three figures, the exact-modulation variance grows from 0.02 to 449 across the sweep, and the normalized variant stays bounded in $[0.045, 0.53] \leq \frac{3}{2}$, as the proposition predicts. The normalized estimator is therefore the stable choice for large-radius or large-bias data, at the cost of estimating the rescaled kernel.

Table 12: Flat-estimator variance of exact vs. normalized modulation ($\varepsilon = 1$, 2×10^5 draws). The ratio equals $(R^2 + b)^4$ exactly; exact modulation blows up while the normalized estimator stays bounded ($\leq \frac{3}{2}$).

b	b -sweep ($R=1$)				R -sweep ($b=1$)				
	exact	norm.	ratio	$(1+b)^4$	R	exact	norm.	ratio	$(R^2+1)^4$
0	0.02	0.02	1.0	1	0.5	0.93	0.38	2.4	2.4
1	3.39	0.21	16	16	1.0	3.38	0.21	16	16
2	29.3	0.36	81	81	2.0	45.6	0.07	625	625
4	331	0.53	625	625	3.0	449	0.045	10^4	10^4

Fixed-dataset radial sample complexity. The radial sample count needed for a target accuracy grows little with dimension (and then plateaus) whereas Nyström’s landmark count does not stay bounded at all. With the recommended flat $D' = 1$ (so D is the number of cosine draws), for each $d \in \{2, \dots, 100\}$ ($N = 500$ on the sphere, $b = \varepsilon = 1, 3$ seeds) we find D^* , the smallest D reaching relative Frobenius error ≤ 0.10 , and the analogous Nyström count m^* (Table 13, Figure 4b).

Table 13: Radial samples needed for relative Frobenius error ≤ 0.10 vs. dimension ($b = 1$, $\varepsilon = 1$, unit sphere, flat $D' = 1$). RAY’s radial count D^* stays bounded and plateaus; Nyström’s landmark count m^* exceeds the tested range (> 350) for $d \geq 20$. “ > 350 ” means 350 landmarks did not reach the target.

d	2	5	10	20	50	100
RAY D^*	200	350	500	750	750	750
Nyström m^*	20	100	350	> 350	> 350	> 350

RAY’s D^* rises from 200 to 750 and then plateaus (no further growth from $d = 20$ to 100), while Nyström reaches the target only for $d \leq 10$: from $d = 20$ on, 350 landmarks leave the error above 0.10 (at $d = 100$, 0.20). The contrast is between a radial count that saturates in d and Nyström’s curse of dimensionality $O(m^{-2s/d})$, the empirical content of Table 2; the total RAY feature dimension is D^*d_b , which does grow with d through the polynomial factor.

Downstream kernel ridge regression. The exact polynomial factor becomes a large downstream advantage at a low number of random draws. On two real datasets (**digits** (classification, $d = 64$) and **california** (regression, $d = 8$), standardized and ℓ_2 -normalized to the sphere, with ε the median squared distance, $b = 1$, and ridge $\lambda = 10^{-2}$, all fixed rather than tuned on validation) we fit kernel ridge regression in Gram form (to isolate kernel-approximation quality; Appendix D.4 evaluates primal-feature scalability separately) and compare, at a *matched number of random draws* D , RAY against IMQ random features (the radial-only estimator of Section 3.4), Gaussian RFF, and Nyström on the exact \mathbf{E} -kernel, with the three exact kernels as references. Matching D matches the Gram-assembly compute up to the one-time $O(N^2d)$ polynomial Gram, a small overhead next to the gap below. Figure 6 reports the mean over 3 splits with ± 1 standard-deviation bands.

With only $D = 32$ radial samples RAY already reaches 0.980 accuracy on digits and comes within 0.01 RMSE of the exact \mathbf{E} -kernel on california, while Gaussian RFF, IMQ-RFF, and Nyström at the same budget trail by a wide margin (0.81–0.89 on digits): they spend their draws approximating the whole kernel, whereas RAY keeps the alignment numerator exact and spends its D draws only on the radial factor. The IMQ-RFF ablation isolates this: removing the numerator and adding it back lifts digits accuracy from 0.81 to 0.98 at $D = 32$, so the exact numerator is the source of the low-budget efficiency. As an *exact* kernel the \mathbf{E} -kernel is

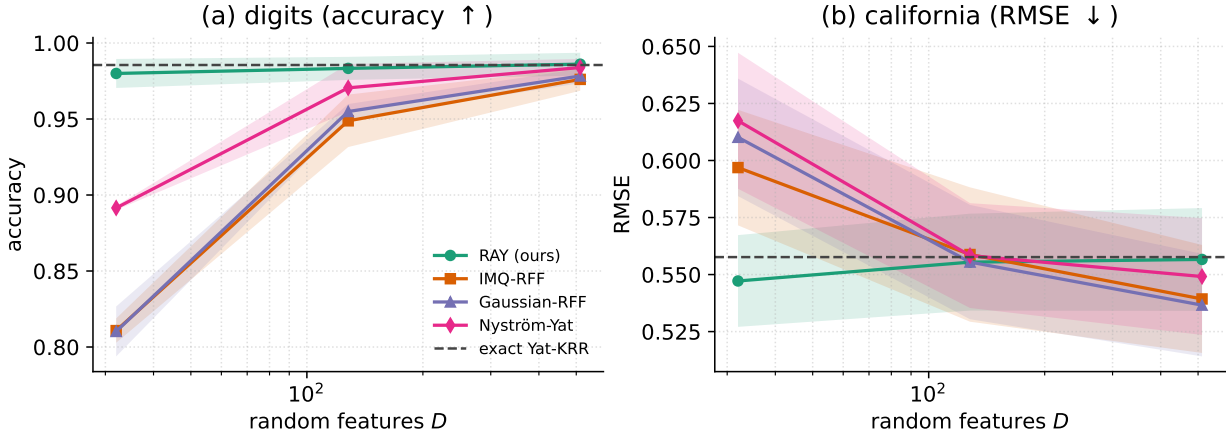


Figure 6: Downstream KRR test metric vs. the number of random draws D on sphere-normalized real data (mean over 3 splits, ± 1 s.d. bands); the dashed line is the exact \mathbf{E} -kernel. **(a)** digits: RAY- \mathbf{E} sits at the exact-kernel accuracy already at $D = 8$, while Gaussian RFF, IMQ-RFF, and Nyström climb slowly and need ~ 512 features to catch up. **(b)** california: RAY tracks the exact kernel from the smallest budgets. RAY keeps the alignment numerator exact and spends its D draws only on the radial factor; the others must sample the whole kernel. As *exact* kernels the \mathbf{E} -kernel ties Gaussian on digits (0.986) and trails it on california (0.558 vs. 0.531 RMSE, IMQ 0.535).

competitive but not uniformly best (it ties Gaussian on digits and trails it by 0.03 RMSE on california); the gain here is RAY’s draw efficiency. Two caveats define the scope: the comparison is matched in the number of *draws*, not in explicit feature dimension, which for RAY carries the polynomial factor $d_b = O(d^2)$ per draw, so the draw-efficiency advantage does not transfer to a dimension-matched setting in high d , where the on-sphere dot-product route (a direction we leave to future work) is preferable; and the \mathbf{E} -kernel needs inputs of bounded norm: on raw standardized features its unbounded numerator destabilizes both the exact kernel and RAY (exact- \mathbf{E} RMSE 0.84, RAY diverging), so we keep $\|x\|$ bounded (here by sphere normalization; Section 5.1 validates the off-sphere bounded-ball regime).

The relative-spectral KRR theorem (Theorem 4.7) holds as stated. On a digits subset ($N = 600$, $\lambda = 0.1$) the whitened error $\rho = \|A^{-1/2}(K_D - K)A^{-1/2}\|_{\text{op}}$ falls with D (log-log slope -0.83), and once $\rho < 1$ (at $D = 1024$, $\rho = 0.69$) the predicted bound $\|\tilde{\alpha} - \hat{\alpha}\|_A \leq \frac{\rho}{1-\rho} \|\hat{\alpha}\|_A$ holds with a wide margin ($0.18 \leq 2.18$); at smaller D the relative error decreases monotonically ($1.59 \rightarrow 0.18$ over $D:16 \rightarrow 1024$) as the deterministic guarantee predicts once it becomes active.

Real-data check: competitive across regimes. The coupled-target preference study (Section 5.3) is synthetic by construction. The same reading (a single \mathbf{E} -kernel stays competitive whichever geometry carries the labels, rather than failing on one) can be tested on real data. We embed CIFAR-10 and CIFAR-100 with a frozen CLIP image encoder (Radford et al., 2021) and run exact kernel ridge regression with the Gaussian, IMQ, degree-2 polynomial, and \mathbf{E} kernels under two preprocessings of the embeddings: centered, where class identity is angular and alignment dominates, and bounded-ball, where it is radial and proximity dominates (Table 14). Across five seeds the four kernels agree to within about one standard deviation: on CIFAR-10 they are statistically tied at ≈ 0.944 , and on CIFAR-100 the polynomial (alignment) kernel leads by roughly one standard deviation with the \mathbf{E} -kernel close behind. The reading is modest but consistent with the coupling picture: the \mathbf{E} -kernel is competitive in both regimes, never far from the best, because it carries whichever factor is informative, whereas a single-factor kernel would be exposed on the regime it is blind to; on these embeddings, though, one factor already suffices, so no kernel separates decisively. And this is a statement about the *kernel*: at $d = 512$ the random-feature \mathbf{E} needs more draws than a single-factor RFF to realize it, consistent with the $O(d^2)$ representation floor (Limitation (v)).

Table 14: Exact-kernel KRR test accuracy on frozen CLIP embeddings of CIFAR-10/100 (10k train / 2k test, fixed λ), under centered (alignment-dominated) and bounded-ball (proximity-dominated) preprocessing. Means over 5 random train/test subsamples; per-cell standard deviation ≤ 0.006 . On CIFAR-10 all four kernels agree within one standard deviation; on CIFAR-100 the polynomial leads by about one standard deviation. Best mean per row in bold (CIFAR-10 alignment is a four-way tie).

dataset	regime	Gaussian	IMQ	polynomial	\mathbf{E}
CIFAR-10	alignment	0.944	0.944	0.944	0.944
CIFAR-10	proximity	0.942	0.942	0.944	0.942
CIFAR-100	alignment	0.768	0.766	0.771	0.766
CIFAR-100	proximity	0.763	0.760	0.768	0.761

Real coupled-target test: QM9 atomization energy. The coupled-target preference (Section 5.3) is synthetic; QM9 (Ramakrishnan et al., 2014) is the natural real-data test, because atomization energy is the home benchmark of kernel ridge regression in chemistry (Rupp et al., 2012) and is extensive: size and composition live in the *norm* of a Coulomb-matrix descriptor, structure in its *direction*. We run the tuned protocol of Table 9 on all 133,885 molecules (targets U_0 atomization in kcal/mol; train 6000/val 1500/test 2000, 3 resamples; 99th-percentile norm clip then global scaling, so the data is an off-sphere ball) with two descriptors: the Coulomb-matrix eigenspectrum ($d = 29$) and the sorted full Coulomb matrix ($d = 435$), and with the two coupling controls of the necessity test: a *direction-only* ablation (inputs projected to the sphere) and a *norm-only* ablation (1-D input $\|x\|$), plus the composition-linear “dressed-atom” baseline (Table 15).

Table 15: QM9 atomization energy, test MAE in kcal/mol (mean \pm std over 3 resamples; train 6000 unless stated; per-kernel ε -multiplier/ b/λ tuned on validation). Joint = full descriptor; dir-only = sphere-projected (norms deleted). Norm-only (1-D $\|x\|$, Gaussian): 172.9 ± 14.7 (eig) / 161.2 ± 1.6 (full CM), against ≈ 190 for predicting the mean. Dressed atom (linear in element counts): 19.6. Full- N sketched-RAY primal trains on all 131,885 rows ($m=128$; eig: $D=32$, $M=5056$, 12s; full CM: $D=16$, $M=9024$, 35s), where the exact modulation ($d_b \approx 95,000$ at $d=435$) and the $N \times N$ Gram are both impossible.

kernel	eigenspectrum ($d=29$)		full CM ($d=435$)	
	joint	dir-only	joint	dir-only
Gaussian	31.6 \pm .3	35.0 \pm 1.0	16.2 \pm .4	16.6 \pm .4
IMQ	35.6 \pm .8	37.3 \pm .5	15.9 \pm .5	16.2 \pm .4
Matérn- $\frac{1}{2}$	45.7 \pm 1.2	45.3 \pm .5	18.3 \pm .3	18.5 \pm .3
polynomial	32.8 \pm .3	35.5 \pm .6	17.0 \pm .4	17.5 \pm .5
\mathbf{E} (exact)	35.7 \pm .7	37.3 \pm .5	15.9 \pm .5	16.3 \pm .4
RAY ($D=1000$)	32.6 \pm .4	–	16.9 \pm .3	–
full- N RAY primal (131,885 rows)	26.8	–	14.5	–

The verdict is honest in both directions. *The coupling is real but modest:* deleting the norms costs every kernel a consistent 2–11% (Gaussian $31.6 \rightarrow 35.0$ on the eigenspectrum, $16.2 \rightarrow 16.6$ on the full CM), while the norm alone is nearly uninformative ($161\text{--}173$ vs. ≈ 190 for the mean), so the target is direction-dominated with a small genuine norm contribution rather than strongly coupled. *The kernel preference does not transfer:* the \mathbf{E} -kernel is statistically tied with the best kernel on the full Coulomb matrix ($15.9 \pm .5$ vs. IMQ $15.9 \pm .5$, Gaussian $16.2 \pm .4$) and behind the Gaussian on the eigenspectrum, so the synthetic coupled-target advantage of Table 7 does not reappear here. Two findings are positive and specific to the construction. First, RAY *beats its own exact kernel* on the eigenspectrum (32.6 vs. 35.7): the random-feature low-rank truncation acts as a regularizer on a near-noiseless target, an effect the unbiasedness analysis permits but does not predict. Second, and the operative point, the full-dataset sketched primal dominates everything Gram-bound: at $d = 435$ the exact modulation feature ($d_b \approx 95,000$ per draw) cannot be built and the 1.7×10^{10} -entry Gram cannot be formed, yet the $M=9024$ sketched primal trains on all 131,885 molecules in 35s and

reaches 14.5 kcal/mol, beating the best exact kernel (15.9) that is confined to a 6000-point Gram: more data through compressed features beats a better-resolved kernel on a subsample. Scripts `qm9_build_cache`, `qm9_atomization`.

Importance sampling (Section 4.3). The proposal $t \sim \text{Exp}(\varepsilon + \eta)$ is unbiased but has a finite-variance window.

Proposition D.3 (Exponential proposal for the radial factor). *With $T \sim \text{Exp}(\varepsilon + \eta)$ ($\eta \geq 0$) and $\hat{h}_\eta(r) = \frac{1}{\varepsilon + \eta} e^{-(r-\eta)T}$, we have $\mathbb{E}[\hat{h}_\eta(r)] = (r + \varepsilon)^{-1}$, and $\mathbb{E}[\hat{h}_\eta(r)^2] = ((\varepsilon + \eta)(\varepsilon + 2r - \eta))^{-1}$ is finite iff $\eta < \varepsilon + 2r$. A proposal safe for all pairs, including $r = 0$, requires $\eta < \varepsilon$.*

Proof. $\mathbb{E}[\hat{h}_\eta(r)] = \int_0^\infty \frac{1}{\varepsilon + \eta} e^{-(r-\eta)t} (\varepsilon + \eta) e^{-(\varepsilon + \eta)t} dt = \int_0^\infty e^{-(r+\varepsilon)t} dt = (r + \varepsilon)^{-1}$. Similarly $\mathbb{E}[\hat{h}_\eta(r)^2] = \frac{1}{\varepsilon + \eta} \int_0^\infty e^{-(\varepsilon + 2r - \eta)t} dt$, finite iff $\varepsilon + 2r - \eta > 0$, equal to $((\varepsilon + \eta)(\varepsilon + 2r - \eta))^{-1}$. \square

Empirically, on pairs at several squared distances ($x^\top w = 0.5$, $b = 1$, $\varepsilon = 1$, $\eta < \varepsilon$, $D = 200$, 3000 repetitions) the estimator stays unbiased (bias $\leq 10^{-3}$) and the reduction is real but tuning-dependent: up to a $0.42\times$ variance ratio at $\|x - w\|^2 = 1$ ($\eta = 0.5$) and $0.55\times$ at $\|x - w\|^2 = 0.25$ ($\eta = 0.1$), but *no* improvement for the nearest pair $\|x - w\|^2 = 0.05$ (where $\eta = 0$ is best). Since the optimal η is data-dependent, a single fixed proposal is a compromise.

Orthogonal features in $d > 1$ (Section 4.3). At a fixed scale t we estimate the Gaussian factor $g_t(x, w)$ with $D' = d$ inner frequencies, drawn either i.i.d. $\mathcal{N}(0, 2tI_d)$ or as a scaled Haar-orthogonal block (Yu et al., 2016) that preserves the marginal. Both are unbiased (empirical bias $\leq 6 \times 10^{-3}$), and the orthogonal block lowers the inner variance by a ratio of 0.72 at $d = 5$ and 0.76 at $d = 20$ (4000 repetitions, $x^\top w = 0.3$). Unlike the $d = 1$ case of Appendix D.2, where there is no direction to decorrelate, orthogonality helps once $d > 1$.

Polynomial sketching (Section 6). The TensorSketch that RAY uses (Proposition A.9) sketches only the quadratic term $(x^\top w)^2$ and keeps $2bx^\top w + b^2$ exact. Here we probe the cruder variant that sketches the whole augmented feature (x, \sqrt{b}) , replacing the exact d_b -dimensional polynomial by a D_{poly} -dimensional sketch (this also sketches the bias and linear terms, so we do not use it in the main experiments). Its error decays slowly, as $\sim D_{\text{poly}}^{-1/2}$: at $d = 10$ ($d_b = 121$) a sketch of $D_{\text{poly}} = 128$ gives full-Gram error 0.175 against the exact feature’s 0.046; at $d = 100$ ($d_b = 10,201$) a $20\times$ smaller sketch ($D_{\text{poly}} = 512$) gives 0.172 against 0.060. Sketching is thus a controllable memory–accuracy trade, useful only when d^2 is prohibitive and some accuracy can be sacrificed; at moderate d the exact polynomial feature is both cheaper and more accurate. The d^2 cost is better reduced losslessly by the symmetric $d(d+1)/2$ reduction.

A non- \mathbf{E} Bernstein–Schur instance (Theorem 3.6). To check that the construction is genuinely about the class and not the \mathbf{E} -kernel alone, we run it unchanged on a different member: $k(x, w) = (x^\top w + b)^3 (\|x - w\|^2 + \varepsilon)^{-\alpha}$, a degree-3 polynomial modulation times the generalized IMQ ($\alpha = 2$), whose Bernstein measure is $\Gamma(\alpha, \varepsilon)$ with shape α and rate ε . Keeping the cubic modulation exact and sampling $T \sim \Gamma(\alpha, \varepsilon)$, $\omega \sim \mathcal{N}(0, 2TI_d)$, on off-sphere data ($\|x\| \in [0.3, 1.2]$, $d = 8$, $N = 400$) the estimator is unbiased (mean of 200 estimates at $D = 50$ is within 0.023 relative Frobenius error of the exact Gram) and converges at the Monte-Carlo rate (relative error $0.71 \rightarrow 0.078$ as D runs $10 \rightarrow 1000$, fitted log-log slope -0.48). Only the modulation feature and the mixing law changed from the \mathbf{E} -kernel; the recipe did not.

A kernel-grammar instance on real data (Theorem 4.13). The compositional-search product $\text{LIN}^2 \times \text{RQ}$ (Duvenaud et al., 2013) is exercised end to end on California housing ($N = 20,640$, $d = 8$; standardized, 99th-percentile norm clip so the ball is off-sphere, targets standardized). With $k(x, w) = (x^\top w + b)^2 (1 + \|x - w\|^2 / \varepsilon')^{-\alpha}$, $b = 1$, $\alpha = 2$, $\varepsilon' = 2\alpha\sigma^2$ set to the median squared distance (0.225), the class estimator (modulation exact, $T \sim \Gamma(\alpha, \varepsilon')$, trig RFF) gives: (a) Gram fidelity at $N = 2000$ following the Monte-Carlo rate (relative Frobenius error $0.31 \rightarrow 0.15 \rightarrow 0.089 \rightarrow 0.041$ over $D = 50/200/800/3200$, fitted slope -0.47 ; 3 seeds); (b) downstream KRR at $N = 3000$ matching the exact composite kernel already at $D = 200$ (test RMSE $0.540 \pm .007$ vs. exact 0.543, $\lambda = 10^{-2}$); (c) full-scale primal ridge on all $N = 19,640$ training rows with explicit features ($D = 100$, $d_b = 45$, $M = 4500$) fitting in 1.2s at test RMSE 0.480, *below* the

exact-kernel reference because the reference is Gram-bound to a 3000-point subsample while the primal uses all the data; that gap is the scaling argument in one number. Script `grammar_kernel`.

Sensitivity to b and ε . We sweep the bias $b \in \{0, 0.1, 1, 10\}$ and the radial scale $\varepsilon \in \{0.25, 1, 4\} \times$ (median squared distance) and report the relative Frobenius Gram error of the flat estimator ($D = 200$, $N = 300$, $d = 16$, 5 seeds) for exact, normalized, and sketched modulation (the last is RAY) (Table 16). The exact and normalized estimators stay in a tight 0.06–0.25 band across two orders of magnitude in b and a $16\times$ range in ε : the method is insensitive to either, and $\varepsilon =$ median squared distance is a sound default. RAY adds the expected $O(m^{-1/2})$ sketch error ($m = 128$), largest at small b and small ε where the quadratic term dominates the modulation.

Table 16: Relative Frobenius Gram error vs. the bias b and radial scale ε (as a multiple of the median squared distance), flat estimator at $D = 200$ ($N = 300$, $d = 16$, mean over 5 seeds; per-cell standard deviation ≤ 0.023 for the exact and normalized estimators, ≤ 0.046 for RAY (sketch)). The exact and normalized estimators are stable across the grid; RAY adds an $O(m^{-1/2})$ sketch error, largest where the quadratic modulation dominates (small b , small ε).

b	$\varepsilon = 0.25\times$			$\varepsilon = 1\times$			$\varepsilon = 4\times$		
	exact	norm	RAY	exact	norm	RAY	exact	norm	RAY
0	0.095	0.114	0.256	0.081	0.083	0.384	0.063	0.062	0.547
0.1	0.092	0.107	0.231	0.077	0.073	0.337	0.058	0.057	0.481
1	0.190	0.165	0.201	0.102	0.086	0.123	0.065	0.061	0.106
10	0.251	0.235	0.251	0.111	0.107	0.111	0.069	0.067	0.069

D.2 Kernel-value variance and variance reduction

The estimator’s variance falls as $O(1/D)$, and quasi-Monte Carlo over the scale lowers it by a constant factor. We estimate $k_{\mathbb{E},b}(x, w)$ for a fixed $d = 1$ pair over 1000 repetitions ($x = 0.5$, $w = 1$, $b = \varepsilon = 1$, $D' = 50$; the scalar exact polynomial factor isolates the radial Gaussian approximation), comparing the basic, QMC, orthogonal, and combined estimators in Table 17.

Table 17: Empirical variance of the biased-kernel estimator ($\times 10^{-3}$, $b = 1$, $\varepsilon = 1$, 1000 repetitions).

Method	$D = 10$	$D = 50$	$D = 100$
Basic RAY	17.54	3.92	2.00
QMC RAY	7.32	1.23	0.62
Orthogonal RAY	19.69	4.28	2.21
QMC + orthogonal	8.31	1.40	0.69

Basic RAY halves its variance with each doubling of D (a $4.5\times$ drop from $D=10$ to 50), and QMC sampling of the radial scale cuts it a further 2.4–3.2 \times at every D . Orthogonalizing the inner frequencies does nothing here (in one dimension there is no direction to decorrelate) but reduces the inner variance by 24–28% once $d > 1$ (Appendix D) and composes with QMC. The Theorem A.1 bound holds but is loose, its worst-case Cauchy–Schwarz making the empirical constants far smaller (at $D=100$, bound 0.16 vs. variance 0.002).

One caveat on QMC: it buys a constant, not a faster rate. Fitting log-log RMSE against $D \in [10, 2000]$ gives a Monte-Carlo slope of -0.502 and a QMC slope of -0.514 , the same $O(1/\sqrt{D})$ rate, with QMC a constant $\sim 1.7\times$ lower. The inner Gaussian RFF noise, which is not quasi-randomized, sets the convergence, so quasi-randomizing only the one-dimensional scale cannot change it; the $O((\log D)^s/D)$ rate of Section 4.3 is the outer integral’s alone.

D.3 Outer/inner allocation: flat sampling is optimal

Flat sampling (one frequency per scale) is the best use of a fixed budget, confirming Proposition A.3. The estimator splits its randomness between D outer scales and D' inner frequencies at identical cost $O(N^2 DD')$; holding the budget $B = DD' = 1000$ fixed and varying the split ($d = 10$, $N = 400$, sphere, $\varepsilon = b = 1$, 4 seeds) gives Table 18.

Table 18: Relative Frobenius error at a fixed budget $B = DD' = 1000$ as the inner count D' varies (so $D = 1000/D'$). Smaller D' is better; $D' = 1$ (flat sampling) wins.

D' (with $D = 1000/D'$)	1	2	5	10	50
rel. Frobenius error	0.059	0.063	0.066	0.089	0.175

Error rises monotonically with D' : $D' = 1$ is optimal here, $3\times$ better than $D' = 50$. This is Proposition A.3 in numbers: with DD' fixed the inner term $V_{\text{in}}/(DD')$ is constant and the outer term V_{out}/D falls as $1/D$ when $V_{\text{out}} > 0$, so the largest D wins (if $V_{\text{out}} = 0$, all allocations tie). The two-level scheme is thus an artifact away from that degenerate case: one should draw D independent (t_j, ω_j) pairs, the flat estimator (8) that Section 3.4 identifies with IMQ random features. (Accordingly, the main Gram, dimension, and ridge-regression experiments all use the recommended flat $D' = 1$; only the variance-reduction probe of Appendix D.2, which isolates the radial Gaussian factor, fixes $D' = 50$.)

D.4 Scalability: clearing the $O(N^2)$ wall

Random \mathbf{E} -features clear the $O(N^2)$ bottleneck the construction was built to clear. We time fitting ridge regression on random targets (cost, not accuracy) three ways (exact kernel ridge (form K , solve the $N \times N$ system); RAY primal ridge (features $Z \in \mathbb{R}^{N \times M}$ with $M = D d'_b$, the symmetric polynomial $d'_b = d(d+1)/2 + d + 1$ and flat $D' = 1$, solve the $M \times M$ system); and Nyström (m landmarks)) reporting wall-clock and the dominant array’s memory ($d = 8$, $D = m = 64$, $M = 2880$; Figure 7).

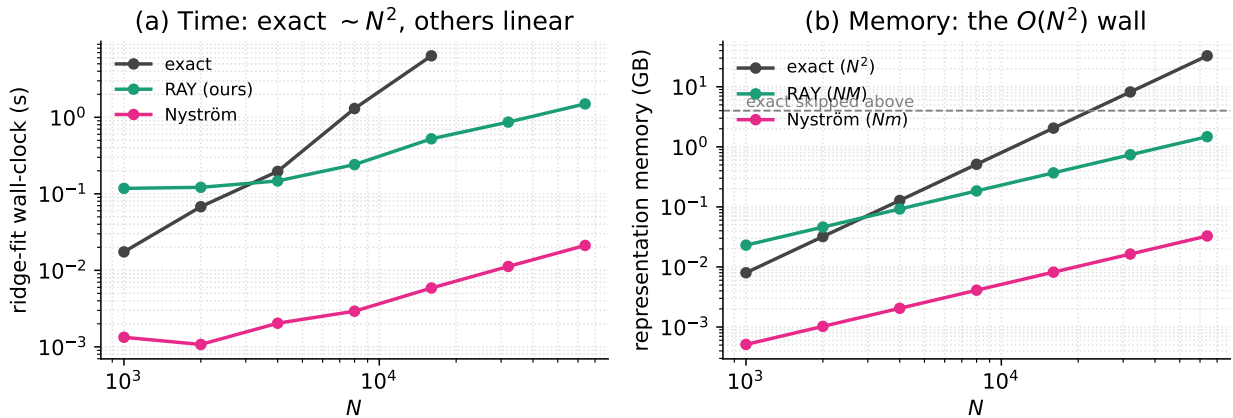


Figure 7: Cost of fitting ridge regression vs. N ($d = 8$, $D = m = 64$, log-log). (a) Wall-clock: exact ridge steepens at the predicted $\sim N^2$ rate (fitted exponent 2.1) and is run only while feasible; RAY and Nyström grow gently. (b) Representation memory: the exact $N \times N$ Gram reaches 33 GB by $N = 64,000$ (above the dashed cap, where it no longer fits), while RAY (NM) and Nyström (Nm) stay linear in N .

Exact ridge scales as the predicted $O(N^2)$ (fitted time exponent 2.1) and its Gram matrix reaches 33 GB at $N = 64,000$, where it no longer fits in memory; RAY and Nyström, whose representations are linear in N (NM and Nm), run there in 1.5 and 0.02 seconds. The honest caveat is the *constant*, not the scaling: RAY’s representation $M = D d'_b$ carries the degree-2 polynomial dimension, so it is heavier than Nyström’s

m (2880 vs 64 here); the symmetric reduction (used) and sketching shrink d'_b , but a downstream method built on RAY should keep D and the polynomial dimension modest.

Large-scale primal training on HIGGS (full sweep). The streaming HIGGS demonstration of Section 5.4 compares, at matched representation dimension M , Gaussian RFF ($D=M$), exact modulation ($D=\lfloor M/d_b \rfloor$), and RAY ($m=128, D=\lfloor M/(m+d+1) \rfloor$); the off-sphere preprocessing standardizes then rescales by the 99.9th norm percentile (varying norms in the bounded ball), ε is the median squared distance, $b=1$, and the polynomial floor is $d_b=435$. Two facts come through (Table 19). First, the estimator trains at 11M scale as an ordinary streaming primal model: peak memory is flat at 8.5 GB, identical to plain random features, because the Gram is never formed. Second, at matched representation budget compression matches or beats exact modulation everywhere and wins decisively at the top budget ($0.751\pm.002$ vs. $0.740\pm.003$ at $M=8192$): the exact polynomial starves the radial factor to $D=1$ at $M=512$, where its AUC swings 0.659–0.724 across seeds (± 0.027), an instability that is itself the argument for compression, while the sketch buys 3–52 \times more draws and stable estimates (± 0.001 – $.009$). Gaussian RFF leads on AUC, which is expected, HIGGS is a smooth detector-feature classification with no strong alignment \times proximity coupling for the \mathbf{E} -modulation to exploit. The claim is not “RAY wins HIGGS” but that compressed Bernstein–Schur features are a practical, memory-flat, million-scale estimator, and that compression, not exactness, is the right setting once d_b is large.

Table 19: Streaming primal training on HIGGS ($N_{\text{train}}=10.5\text{M}$, $N_{\text{test}}=500\text{k}$, $d=28$, $d_b=435$) at matched representation dimension M . Test AUC (\uparrow), mean \pm std over 3 feature seeds; D = radial draws afforded at that budget; best feature method per M in bold. Peak memory is flat at 8.5 GB across all methods (no Gram, no full feature matrix); one sweep runs in ~ 5 minutes on one M5 Pro GPU. Linear baseline AUC 0.682. The exact-modulation column at $M=512$ is a *single* radial draw, whose ± 0.027 seed spread is itself the instability that compression removes.

M	Gaussian RFF		exact modulation		RAY (sketch)	
	D	AUC	D	AUC	D	AUC
512	512	0.721 $\pm.002$	1	0.693 $\pm.027$	3	0.697 $\pm.009$
1024	1024	0.738 $\pm.001$	2	0.714 $\pm.012$	6	0.718 $\pm.001$
2048	2048	0.749 $\pm.001$	4	0.730 $\pm.006$	13	0.731 $\pm.002$
4096	4096	0.755 $\pm.001$	9	0.737 $\pm.004$	26	0.742 $\pm.001$
8192	8192	0.759 $\pm.000$	18	0.740 $\pm.003$	52	0.751 $\pm.002$

Protocol. Streaming Adam (learning rate 0.02, batch size 8192, two passes over the 10.5M training rows), single precision (fp32), three feature seeds (mean \pm std in the table); features are rebuilt per mini-batch on the GPU and discarded. Reported wall-clock includes per-batch feature construction but excludes the one-time gzip load of the data; the 8.6 GB peak (measured by `getrusage`) is the resident process maximum, the same for all methods since none forms a Gram. The whole $\{512, \dots, 8192\} \times \{\text{linear, Gaussian, exact modulation, RAY}\}$ sweep takes ~ 260 s on one Apple M5 Pro.

Relative-spectral KRR stability for the sketched estimator. The relative-spectral guarantee (Theorem 4.7) also governs the deployed sketched RAY (Table 20). The sketch term keeps ρ above the exact-modulation value, so more regularization or budget is needed for the bound to activate, but once it does the coefficient error obeys it.

Bounded-input preprocessing. The unbounded numerator of $k_{\mathbf{E},b}$ makes a norm bound mandatory (Limitation (iii), Table 21). Raw standardization (max norm 9.3) inflates RMSE to 0.37 (exact) / 0.96 (RAY); any bounded scheme recovers ≤ 0.05 , confirming Limitation (iii).

Sphere-normalized matched-cost checks. On the unit sphere $k_{\mathbf{E},b}$ coincides with a dot-product kernel, where direct zonal/dot-product routes are natural baselines; the sphere-normalized matched-dimension and fair-cost results below are therefore sanity checks, with the off-sphere versions (Tables 5, 8) the primary

Table 20: Relative-spectral KRR stability for the deployed (sketched) RAY (digits subset, $N=600$). $\rho = \|A^{-1/2}(K_{D,m} - K)A^{-1/2}\|_{\text{op}}$ with $A = K + \lambda I$; the deterministic bound $\|\tilde{\alpha} - \hat{\alpha}\|_A \leq \frac{\rho}{1-\rho} \|\hat{\alpha}\|_A$ activates once $\rho < 1$. ρ falls with the radial draws D , the sketch size m , and the ridge λ ; the measured coefficient error stays within the bound throughout.

sweep	ρ	rel. coef. err	bound $\rho/(1-\rho)$
λ sweep ($D=512, m=256$)			
$\lambda=0.3$	0.757	0.255	3.12
$\lambda=1.0$	0.517	0.192	1.07
$\lambda=3.0$	0.330	0.138	0.49
$\lambda=10$	0.177	0.088	0.22
D sweep ($m=256, \lambda=3$)			
$D=64$	0.990	0.259	103
$D=256$	0.375	0.158	0.60
$D=1024$	0.292	0.124	0.41
m sweep ($D=512, \lambda=3$)			
$m=64$	0.821	0.219	4.58
$m=128$	0.518	0.173	1.07
$m=512$	0.248	0.108	0.33

Table 21: Bounded-input preprocessing (off-sphere coupled target, $d=16$, KRR test RMSE, mean over 3 seeds). The unbounded numerator destabilizes both the exact kernel and RAY on raw data; any norm bound restores accuracy; clipping and max-norm scaling are best.

scheme	max $\ x\ $	exact \mathbf{E} (RMSE)	RAY (RMSE)
raw (standardize only)	9.25	0.373	0.959
max-norm scaling	1.00	0.029	0.033
percentile clip (99)	1.00	0.027	0.041
sphere normalization	1.00	0.035	0.047
normalized kernel	1.00	0.041	–

evidence. The ordering is the same: at matched dimension RAY’s TensorSketch recovers most of the efficiency the exact feature loses to its $O(d^2)$ size (Table 22), and at matched representation the adaptive Nyström variants are most accurate while RAY is the data-independent streaming option (Table 23).

Table 22: Matched-dimension KRR on sphere-normalized `digits` ($d = 64, b = 1$, accuracy, mean \pm std over 3 splits). At a fixed explicit feature dimension M , exact modulation affords only M/d_b radial draws; RAY ($m = 128$) compresses the polynomial factor and affords many more, approaching the optimal rank- M ceiling.

M	exact modulation	RAY ($m=128$)	optimal rank- M
2145	0.928 \pm .040	0.977 \pm .006	0.986 \pm .007
4290	0.924 \pm .054	0.977 \pm .006	0.986 \pm .007
8580	0.973 \pm .004	0.979 \pm .007	0.986 \pm .007

Table 23: Fair-cost comparison on sphere-normalized **digits** ($d = 64$, $b = 1$, accuracy mean \pm std over 3 splits) at matched representation dimension. Memory is $N \times \dim$ floats; build is feature/landmark construction wall-clock.

Method	dim	accuracy	memory (MB)	build (s)
exact modulation	2145	0.928 \pm .040	20.6	0.020
RAY ($m=128$)	2123	0.977 \pm .006	20.3	0.009
k-means Nyström	1198	0.986 \pm .007	11.5	1.18
rls-leverage Nyström	1198	0.986 \pm .004	11.5	0.434
Gaussian RFF (different kernel)	2145	0.984 \pm .004	20.6	0.024

E Experimental Details

This appendix collects the protocol shared across experiments and the per-table configuration, so that every number in the paper can be reproduced. The accompanying code, figures, and a project page are available online.²

Shared protocol. Unless a table states otherwise, the radial scale ε is set to the median squared pairwise distance of the inputs (computed on a subsample), the bias is $b = 1$, and the deployed RAY map is the lower-variance quadratic-only TensorSketch (sketch size m , D radial draws, flat $D' = 1$): the degree-2 term is sketched to dimension m and the linear/constant terms are kept exact, so the feature dimension is $D(m+d+1)$. Kernel ridge regression is solved exactly (in the primal (features) for RAY and RFF, and in the dual ($N \times N$) for the exact **E**-kernel and Nyström) with the ridge λ listed per table; “exact modulation” is Definition 3.3 (the $m \rightarrow \infty$ limit). Off-sphere data is a bounded ball with norms drawn over a stated interval (e.g. $\|x\| \in [0.3, 1.5]$), the regime where $k_{\mathbf{E},b}$ is genuinely non-dot-product; sphere-normalized runs (Appendix tables) fix $R = 1$ and are sanity checks only. Reported error bars are mean \pm std over the seed count in each caption.

Hardware and precision. All Gram-error, KRR, operator-norm, and feature-build timing experiments run on CPU in double precision (fp64) and use no GPU. The sole GPU experiment is the million-scale HIGGS streaming run (Table 19), which runs in single precision (fp32) on one Apple M5 Pro GPU; its full protocol (streaming Adam, batch 8192, two passes over the training rows, peak resident memory) is stated inline with the table. Timing tables therefore compare like for like within a row; absolute wall-clock is hardware-specific and reported only to show scaling, not peak throughput.

²<https://www.tahabouhsine.com/ray>

Table 24: Per-table configuration. “Data” gives the dataset and input dimension d ; “train/test” the KRR split sizes (“Gram” or “op-norm” marks experiments that approximate a matrix rather than fit a predictor, with no split; “build” marks feature-construction timing only). Grids list the swept hyperparameters; ε is the median squared distance throughout. Each table’s own caption restates its seed count and error bars.

Table / Fig.	Data (d)	train/test	seeds	Swept grid
Fig. 1, Tab. 4	off-sphere ball ($d \in \{2, 8, 16, 32\}$)	Gram, $N=1000$	5	$D \in \{10, 50, 100, 500, 1000\}$, Nyström $m=100$
Tab. 8	off-sphere ball ($d=64$)	1500/1000	3	$M=2145$, $m_{\text{TS}}=128$, $\lambda=10^{-2}$
Tab. 7	off-sphere ($d=16$)	1333/667	5	$D=4000$, $\lambda=10^{-2}$
Tab. 9	coupled ($d=16$), digits ($d=64$)	800/400 (val 300)	3	$b \in \{0, .5, 1, 2\}$, $\varepsilon_{\text{mult}} \in \{.25, 1, 4\}$, $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}\}$
Tab. 10	off-sphere ($d=32$)	1300/600	3	$M=4096$, $m=128$, $\lambda=10^{-2}$
Tab. 5	off-sphere ($d \in \{16, 64, 256\}$)	400/200	3	$m \in \{16, \dots, 512\}$, $M \in \{4096, 8192, 16384\}$
Tab. 6	synthetic ($d \in \{8, \dots, 1024\}$)	build, $N=1000$	1	$D=8$, $m=128$
Tab. 20	digits ($d=64$)	op-norm, $N=600$	3	$\lambda \in \{.1, .3, 1, 3, 10\}$, $D \in \{64, 256, 512, 1024\}$, $m \in \{64, 128, 256, 512\}$
Tab. 21	off-sphere ($d=16$)	800/400	3	5 norm schemes; $D=24$, $m=128$, $\lambda=10^{-2}$
Tab. 19	HIGGS ($d=28$)	10.5M/500k	3	$M \in \{512, \dots, 8192\}$, $m=128$ (MLX, fp32)
Tab. 22, Tab. 23	sphere digits ($d=64$)	3 splits	3	draws $\in \{1, 2, 4\}$, sketch $\in \{128, 256\}$, $\lambda=10^{-2}$
Fig. 2	synthetic ($d=16$)	op-norm, $N=300$	5	$D \in \{10, 30, 100, 300, 1000\}$, $m \in \{64, 128, 256, 512\}$
Thm. 4.8, Thm. 4.13	synthetic ($d=20$)	whitened op-norm, $N=300$	6	$\lambda \in \{10, 1, .1, .01\}$, $D \in \{25, \dots, 800\}$
Fig. 3	random Q/K/V ($d=32$)	op-fidelity, $N \leq 131072$	3	M up to 12416, $m=64$
Thm. 4.10	synthetic ($d=20$)	whitened op-norm, $N=300$	4	$\lambda \in \{10, 1, .1, .01\}$, $D \in \{25, \dots, 3200\}$, pool 5×10^4
Prop. D.2	off-sphere ($d=16$)	Gram + pointwise, $N=400$	10	$\varepsilon_{\text{mult}} \in \{.25, 1, 4\}$, $D \in \{128, 512, 2048\}$, reps $\leq 10^6$
Rmk. 3.8	off-sphere ($d=16$)	op-norm, $N=300$	10	$m \in \{64, 128, 256, 512\}$, real vs. complex signs
Thm. 4.13 (grammar)	california ($d=8$)	3000/1000 + full primal	3	$D \in \{50, \dots, 4000\}$, $\alpha=2$, $\lambda=10^{-2}$
Tab. 15	QM9 CM eig ($d=29$) / full ($d=435$)	6000/1500/2000 + full primal	3	$\varepsilon_{\text{mult}} \in \{.25, 1, 4\}$, $b \in \{0, .25, 1, 4\}$, $\lambda_{\text{rel}} \in \{10^{-9}, \dots, 10^{-3}\}$