Relative Error Embeddings of the Gaussian Kernel Distance

Di Chen[*] DCHENAD@CONNECT.UST.HK
Noah's Ark Lab, Huawei Technologies
Units 525-530, Core Building 2
Hong Kong Science Park, Shatin, Hong Kong

Jeff M. Phillips[†] JEFFP@CS.UTAH.EDU
University of Utah
50 S Central Campus Dr. 3190
Salt Lake City, UT 84112
United States of America

Editors: Steve Hanneke and Lev Reyzin

Abstract

A reproducing kernel defines an embedding of a data point into an infinite dimensional reproducing kernel Hilbert space (RKHS). The norm in this space describes a distance, which we call the kernel distance. The random Fourier features (of Rahimi and Recht) describe an oblivious approximate mapping into finite dimensional Euclidean space that behaves similar to the RKHS. We show in this paper that for the Gaussian kernel the Euclidean norm between these mapped to features has $(1+\varepsilon)$-relative error with respect to the kernel distance. When there are n data points, we show that $O((1/\varepsilon^2)\log n)$ dimensions of the approximate feature space are sufficient and necessary. Without a bound on n, but when the original points lie in $\mathbb{R}^d$ and have diameter bounded by $\mathcal{M}$, then we show that $O((d/\varepsilon^2)\log \mathcal{M})$ dimensions are sufficient, and that this many are required, up to $\log(1/\varepsilon)$ factors. We empirically confirm that relative error is indeed preserved for kernel PCA using these approximate feature maps.

1. Introduction

The kernel trick in machine learning allows for non-linear analysis of data using many techniques such as PCA and SVM which were originally designed for linear analysis. The "trick" is that these procedures only access data through inner products between data points, and the standard dot product can be replaced with a non-linear inner product kernel $K(\cdot,\cdot)$. Now given n data points, one can compute the $n \times n$ gram matrix G of all pairwise inner products; that is so $G_{i,j} = K(x_i, x_j)$ for all x_i, x_j in input data set X. Then the analysis can proceed using just the gram matrix G.

However, for large data sets, constructing this $n \times n$ matrix is a computational bottleneck, so methods have been devised for lifting n data points $P \subset \mathbb{R}^d$ to a high-dimensional space $\mathbb{R}^m$ (but where $m \ll n$) so that the Euclidean dot product in this space approximates the non-linear inner product defined by K.

For reproducing kernels K, there exists a lifting $\phi : \mathbb{R}^d \to \mathcal{H}_K$, where $\mathcal{H}_K$ is the reproducing kernel Hilbert space. It is in general infinite dimensional, but every finite subset of n points $\Phi(X) = \{\phi(x) \mid x \in X\}$ has the span of an n-dimensional Euclidean space. That is $K(x,y) = \langle \phi(x), \phi(y) \rangle$. Moreover, we can define the norm of a point in $\mathcal{H}_K$ as $\|\phi(x)\|_{\mathcal{H}_K} = \sqrt{\langle \phi(x), \phi(x) \rangle}$ using the inner

[*] Contributed as a student at HKUST while visiting University of Utah, partially supported by RGC grants GRF-16208415, GRF-621413 and GRF-16211614.
[†] Thanks to support by NSF CCF-1350888, IIS-1251019, ACI-1443046, and CNS-1514520.

product, and then due to linearity, a distance (the *kernel distance*) between two points is defined:

$$D_K(x,y) = \|\phi(x) - \phi(y)\|_{\mathcal{H}_K} = \sqrt{\|\phi(x)\|^2_{\mathcal{H}_K} + \|\phi(y)\|^2_{\mathcal{H}_K} - 2\langle \phi(x), \phi(y)\rangle}$$
$$= \sqrt{K(x,x) + K(y,y) - 2K(x,y)}.$$

For reproducing kernels (actually a subset called characteristic kernels) this is a metric Sriperumbudur et al. (2010); Müller (1997).

Thus we may desire an approximate lifting $\hat{\phi} : \mathbb{R}^d \to \mathbb{R}^m$ such that with probability at least $1 - \delta$ for all $x, y \in X$

$$(1-\varepsilon) \leq \frac{D_K(x,y)}{\|\hat{\phi}(x) - \hat{\phi}(y)\|} \leq (1+\varepsilon).$$

It turns out, one can always algorithmically construct such a lifting with $m = O((1/\varepsilon^2)\log(n/\delta))$ by the famous Johnson-Lindenstrauss (JL) Lemma Johnson and Lindenstrauss (1984). However, unlike the JL Lemma, there is not always known an implicit construction. In general, we must first construct the $n \times n$ gram matrix, revealing an n-dimensional subspace (through an $O(n^3)$ time eigendecomposition) and then apply $m = O((1/\varepsilon^2)\log(n/\delta))$ random projections.

So in recent years there have been many types of kernels considered for these implicit embeddings with various sorts of error analysis, such as for Gaussian kernels Rahimi and Recht (2007); Lopez-Paz et al. (2014); Sriperumbudur and Szabo (2015); Sutherland and Schneider (2015) group invariant kernels Li et al. (2010), min/intersection kernels Maji and Berg (2009), dot-product kernels Kar and Karnick (2012), information spaces Abdullah et al. (2016), and polynomial kernels Hamid et al. (2014); Avron et al. (2014).

In this document we reanalyze one of the most widely used and first variants, the Random Fourier Features, introduced by Rahimi and Recht (2007). It applies to symmetric shift-invariant kernels which include Laplace, Cauchy, and most notably Gaussian. We will primarily focus on Gaussian kernels, defined $K(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$, unless specified otherwise. It is characteristic, hence D_K is a metric.

1.1. Existing Properties of Gaussian Kernel Embeddings

Rahimi and Recht (2007) defined two approximate embedding functions: $\tilde{\phi} : \mathbb{R}^d \to \mathbb{R}^m$ and $\hat{\phi} : \mathbb{R}^d \to \mathbb{R}^m$ (defined below). Only the former appears in the final version of paper, but the latter is also commonly used throughout the literature Sutherland and Schneider (2015). Both features use random variables $\omega_i \in \mathbb{R}^d$ drawn uniformly at random from the Fourier transform of the kernel function; in the case of the Gaussian kernel, the Fourier transform is again a Gaussian, specifically $\omega_i \sim \mathcal{N}_d(0, \sigma^{-2})$.

In the former case, they define m functions of the form $\tilde{f}_i(x) = \cos(\langle \omega_i, x\rangle + \gamma_i)$, where $\gamma_i \sim \mathsf{Unif}(0, 2\pi]$, uniformly at random from the interval $(0, 2\pi]$, is a random shift. Applying each $\tilde{f}_i$ on a datapoint x gives the ith coordinate of $\tilde{\phi}(x)$ in $\mathbb{R}^m$ as $\tilde{\phi}(x)_i = \tilde{f}_i(x)/\sqrt{m}$.

In the latter case, they define $t = m/2$ functions of the form

$$\hat{f}_i(x) = \begin{bmatrix} \cos(\langle \omega_i, x\rangle) \\ \sin(\langle \omega_i, x\rangle) \end{bmatrix}$$

as a single 2×1 dimensional vector, and one feature pair. Then applying $\hat{f}_i$ on a data point x yields the $(2i)$th and $(2i+1)$th coordinate of $\hat{\phi}(x)$ in $\mathbb{R}^m$ as $[\hat{\phi}(x)_{2i}; \hat{\phi}(x)_{2i+1}] = \hat{f}_i(x)/\sqrt{t}$.

Rahimi and Recht (2007) showed $\mathbf{E}[\langle \tilde{\phi}(x), \tilde{\phi}(y)\rangle] = K(x,y)$ for any $x, y \in \mathbb{R}^d$, and that this implied

$$\mathbf{Pr}[|\langle \tilde{\phi}(x), \tilde{\phi}(y)\rangle - K(x,y)| \geq \varepsilon] \leq \delta$$

- with $m = O((1/\varepsilon^2)\log(1/\delta))$ for each $x, y \in \mathbb{R}^d$,
- with $m = O((1/\varepsilon^2)\log(n/\delta))$, for all $x, y \in X$, for $X \subset \mathbb{R}^d$ of size n, or
- with $m = O((d/\varepsilon^2)\log(\mathcal{M}/\delta))$, for all $x, y \in X$, for $X \subset \mathbb{R}^d$ so $\mathcal{M} = \max_{x,y\in X} \|x - y\|/\sigma$.

Recently Sriperumbudur and Szabo (2015) improved the constants in these bounds, and showed rate optimality. It is folklore (apparently removed from final version of Rahimi and Recht (2007); reproved Sutherland and Schneider (2015)) that also $\mathbf{E}[\langle\hat{\phi}(x), \hat{\phi}(y)\rangle] = K(x, y)$, and thus all of the above PAC bounds hold for $\hat{\phi}$ as well. Sutherland and Schneider (2015) also compared $\tilde{\phi}$ and $\hat{\phi}$ (they used $\breve{\phi}$ for our $\tilde{\phi}$ and $\tilde{\phi}$ for our $\hat{\phi}$), and demonstrated that $\hat{\phi}$ performs better (for the same m) and has provably lower variance in approximating $K(x, y)$ with $\langle\hat{\phi}(x), \hat{\phi}(y)\rangle$ as opposed to with $\langle\tilde{\phi}(x), \tilde{\phi}(y)\rangle$. However, these results do *not* obtain a bound on $\|\hat{\phi}(x) - \hat{\phi}(y)\|/D_K(x, y)$ since for very small distances, the additive error bounds on $K(x, y)$ are not sufficient to say much about $D_K(x, y)$.

1.2. Our Results

In this paper we show that $\hat{\phi}$ probabilistically induces a kernel $\hat{K}(x, y) = \langle\hat{\phi}(x), \hat{\phi}(y)\rangle$ and a distance

$$D_{\hat{K}}(x, y) = \sqrt{\|\hat{\phi}(x)\|^2 + \|\hat{\phi}(y)\|^2 - 2\hat{K}(x, y)} = \|\hat{\phi}(x) - \hat{\phi}(y)\|,$$

which has strong relative error bounds with respect to $D_K(x, y)$, namely for a parameter $\varepsilon \in (0, 1)$

$$(1 - \varepsilon) \leq \frac{D_K(x, y)}{D_{\hat{K}}(x, y)} \leq (1 + \varepsilon). \tag{1}$$

In Section 2 we show (1) holds for each x, y such that $\|x - y\|/\sigma \geq 1$, with probability at least $1 - \delta$, with $m = O((1/\varepsilon^2)\log(1/\delta))$. We also review basic properties about $\hat{\phi}$ and D_K.

We first prove bounds that depend on the size n of a data set $X \subset \mathbb{R}^d$. We show that $m = O((1/\varepsilon^2)\log n)$ features are necessary (Section 3) and sufficient (Section 4) to achieve (1) with high probability (e.g., at least $1 - 1/n$), when d and X are otherwise unrestricted.

In Section 5 we prove bounds for $X \subset \mathbb{R}^d$ where d is small, but the size $n = |X|$ is unrestricted. Let $\mathcal{M} = \max_{x,y\in X} \|x - y\|/\sigma$. We show that $m = O((d/\varepsilon^2)\log(\frac{d}{\varepsilon}\frac{\mathcal{M}}{\delta}))$ is sufficient to show (1) with probability $1 - \delta$. Then in Section 6 we show that $m = \Omega(\frac{d}{\varepsilon^2\log(1/\varepsilon)}\log(\frac{\mathcal{M}}{\log(1/\varepsilon)}))$ is necessary for any feature map.

In Section 8 we empirically confirm the relative error through simulations. This includes showing kernel PCA obtains relative error using these approximate features.

1.3. Implications in Machine Learning and Data Analysis

These new relative error bounds have numerous implications in machine learning and geometric data analysis. We mention a couple others involving geometric approximations in learning and mining, and in an L_1 bound on Gram matrix approximations in Section 7.

Limits on oblivious kernel embeddings. There has been extensive recent effort to find oblivious subspace embeddings (OSE) of data sets into Euclidian spaces that preserve relative error Avron et al. (2014); Woodruff (2014); Larsen and Nelson (2016); Clarkson and Woodruff (2015). Strong positive results are known for high-dimensional linear kernels (via Johnson-Lindenstrauss Johnson and Lindenstrauss (1984); Woodruff (2014); Larsen and Nelson (2016, 2017)), for polynomial kernels Avron et al. (2014), and for any M-estimator with gradient between 1 and 2 Clarkson and Woodruff (2015), but has remained open for the Gaussian kernel. Such strong guarantees are, for

instance, required to prove results about regression on the resulting set since we may not know the units on different coordinates; additive error bounds do not make sense in directions which are linear combinations of several coordinates.

The obliviousness of the features (they can be defined without seeing the data, and in some cases are independent of the data size) are essential for many large-scale settings such as streaming or distributed computing where we are not able to observe all of the data at once.

Our results do not describe unrestricted OSEs, as are possible with polynomial kernels Avron et al. (2014). Rather our lower bounds show that any OSE must have the dimension depend on n or $\mathcal{M}$.

Kernel k-means clustering. Kernel k-means Girolami (2002) aims to find a set of k center points in $\mathcal{H}_K$ minimizing the sum of squared kernel distances from the $\phi(x) \in \Phi(X)$ to the closest center point.

Typical approaches either use the full $O(n)$-sized representation of the center Girolami (2002) or heuristically approximate $\mathcal{H}_K$ using the top k-eigenvectors of the Gram matrix G (with no individual distance guarantees). In order to perform kernel k-means clustering in the former case, a recurring operation is to invoke the distance computation between the k center points and $\phi(x)$. Due to the representation size of each center point, the operation takes at least time $\Omega(n)$. If an approximate lifting map $\hat{\phi}$ is used instead, the center points can be explicitly represented as a m-dimensional point, and the distance computation would take $O(dm)$ time with bounded relative error. This also means the related sublinear algorithms such as Ailon et al. (2009) can be applied directly, with small space usage, which is not possible if one can only rely on the Gram matrix.

On the other hand, often these methods may use a representative data point $\phi(x) \in \Phi(X)$ instead of the mean of the included data points Dhillon et al. (2004). Then our upper bounds imply one can simply work in Euclidean $\mathbb{R}^m$ space, and have relative error guarantees on the overall cost function found. This still allows us to use spatial indexing or searching techniques such as LSH and k-d trees to speed up algorithms such as k-means++ Arthur and Vassilvitskii (2007) or the Gonzalez (1985) algorithm for kernel k-center clustering.

Kernel distance matching. The kernel distance $D_K(X, Y)$ between two point sets provides a robust and powerful distance between objects X and Y, for instance probability measures Smola et al. (2007); Gretton et al. (2012), medical images of organs Durrleman et al. (2007); Glaunès and Joshi (2006), and general shapes Joshi et al. (2011). However this distance (a single scalar value) *does not* imply or provide an alignment between the point sets (unlike other common integral probability measures, say like the Wasserstein family of distances e.g., earth-movers). Embedding the point sets into $\mathbb{R}^m$, allows one to invoke powerful geometric approaches using Euclidian distance Sharathkumar and Agarwal (2012); Agarwal and Sharathkumar (2014) to construct the *matching* which approximately minimizes the pairwise kernel distance.

2. Basic Bounds and Taylor Approximations

For the remainder of the paper, it will be convenient to let $\Delta = (x - y)/\sigma$ be the scaled vector between some pair of points $x, y \in X$. Define $D_K(\Delta) = D_K(x, y) = \sqrt{2 - 2e^{\frac{1}{2}\|\Delta\|^2}}$, and also $K(\Delta) = K(x, y)$ and $\hat{K}(\Delta) = \hat{K}(x, y)$.

Using $t = O((1/\varepsilon^2)\log(1/\delta))$ features for $\varepsilon \in (0, 1/2)$ and $\delta \in (0, 1)$, then for any $\Delta \in \mathbb{R}^d$, the following PAC bound Sutherland and Schneider (2015); Rahimi and Recht (2007) holds

$$\mathbf{Pr}\left[\left|K(\Delta) - \hat{K}(\Delta)\right| \leq \varepsilon\right] \geq 1 - \delta. \tag{2}$$

Since $\hat{K}(x,x) = 1$, then $D_{\hat{K}}(\Delta)^2 = 2 - 2\hat{K}(\Delta)$, and additive error bounds between $D_K(\Delta)^2$ and $D_{\hat{K}}(\Delta)^2$ follow directly. But we can also state some relative error bounds when $\|\Delta\|$ is large enough.

Lemma 1 *For each $\Delta \in \mathbb{R}^d$ such that $\|\Delta\| \geq 1$ and $m = O((1/\varepsilon^2)\log(1/\delta))$ with $\varepsilon \in (0, 1/10)$ and $\delta \in (0,1)$. Then with probability at least $1-\delta$, we have $\frac{D_K(\Delta)}{D_{\hat{K}}(\Delta)} \in [1-\varepsilon, 1+\varepsilon]$.*

Proof By choosing $m = O((1/\varepsilon^2)\log(1/\delta))$ so that $|K(\Delta) - \hat{K}(\Delta)| \leq \varepsilon/4$, via (2), we have that $|D_K^2(\Delta) - D_{\hat{K}}^2(\Delta)| \leq \varepsilon/2$. We also note that when $\|\Delta\| \geq 1$ then $K(\Delta) \leq \frac{1}{\sqrt{e}} \leq 0.61$. Hence $D_K^2(\Delta) \geq 2(1-0.61) = 0.78 \geq 0.5$, and we have that $|D_K^2(\Delta) - D_{\hat{K}}^2(\Delta)| \leq \varepsilon/2 \leq \varepsilon D_K^2(\Delta)$. Then $|1 - \frac{D_{\hat{K}}^2(\Delta)}{D_K^2(\Delta)}| \leq \varepsilon$, implying $1-\varepsilon \leq \frac{D_{\hat{K}}^2(\Delta)}{D_K^2(\Delta)} \leq 1+\varepsilon$. Taking the square root of all parts completes the proof via $\sqrt{1+\varepsilon} < (1+\varepsilon)$ and $\sqrt{1-\varepsilon} > (1-\varepsilon)$. ■

Basic bounds when $\|\Delta\| < 1$. When $\|\Delta\| \leq 1$, then a simple Taylor expansion, implies that

$$\|\Delta\|^2 - \frac{1}{4}\|\Delta\|^4 \leq D_K(\Delta)^2 = 2 - 2\exp(\|\Delta\|^2/2) \leq \|\Delta\|^2,$$

and by $\frac{1}{4}\|\Delta\|^4 \leq \frac{1}{4}\|\Delta\|^2$ and a square root

$$0.86\|\Delta\| \leq D_K(\Delta) \leq \|\Delta\|. \tag{3}$$

Moreover, when $\|\Delta\| \leq 2\sqrt{\varepsilon}$ then

$$(1-\varepsilon)\|\Delta\|^2 \leq D_K(\Delta)^2 \leq \|\Delta\|^2. \tag{4}$$

Useful expansions. We first observe that by $\cos(a)\cos(b) + \sin(a)\sin(b) = \cos(a-b)$ that

$$\langle \hat{f}_i(x), \hat{f}_i(y)\rangle = \cos(\langle \omega_i, x\rangle)\cos(\langle \omega_i, y\rangle) + \sin(\langle \omega_i, x\rangle)\sin(\langle \omega_i, y\rangle) = \cos(\langle \omega_i, (x-y)\rangle).$$

Hence by $\langle \hat{f}_i(x), \hat{f}_i(x)\rangle = \cos(\langle \omega_i, 0\rangle) = 1$ we have $D_{\hat{K}}(x,y)^2 = 2 - 2\frac{1}{t}\sum_i^t \cos(\langle \omega_i, (x-y)\rangle)$.

By the rotational stability of the Gaussian distribution we can replace $\langle \omega_i, (x-y)\rangle$ with $\omega_{i,x,y}\frac{\|x-y\|}{\sigma}$ where $\omega_{i,x,y} \sim \mathcal{N}(0,1)$. It will be more convenient to write $\omega_{i,x,y}$ as $\omega_{i,\Delta}$, so $\langle \omega_i, (x-y)\rangle = \omega_{i,\Delta}\|\Delta\|$. Thus $\langle \hat{f}_i(x), \hat{f}_i(y)\rangle = \cos(\omega_{i,\Delta}\|\Delta\|)$. Moreover, we can define $D_{\hat{K}}(\Delta) = D_{\hat{K}}(x,y) = \sqrt{2 - 2\frac{1}{t}\sum_{i=1}^t \cos(\omega_{i,\Delta}\|\Delta\|)}$.

Now considering

$$\frac{D_{\hat{K}}(\Delta)^2}{D_K(\Delta)^2} = \frac{1 - \frac{1}{t}\sum_{i=1}^t \cos(\omega_{i,\Delta}\|\Delta\|)}{1 - e^{\frac{1}{2}\|\Delta\|^2}},$$

the following Taylor expansion, for $\omega_{i,\Delta}\|\Delta\| \leq 1$, will be extremely useful:

$$\frac{\frac{1}{t}\sum_{i=1}^t \frac{1}{2}\omega_{i,\Delta}^2\|\Delta\|^2}{\frac{1}{2}\|\Delta\|^2 - \frac{1}{4}\|\Delta\|^4} \geq \frac{D_{\hat{K}}(\Delta)^2}{D_K(\Delta)^2} \geq \frac{\frac{1}{t}\sum_{i=1}^t \left(\frac{1}{2}\omega_{i,\Delta}^2\|\Delta\|^2 - \frac{1}{24}(\omega_{i,\Delta}^4\|\Delta\|^4)\right)}{\frac{1}{2}\|\Delta\|^2}.$$

Simplifying gives

$$\frac{1}{1 - \frac{1}{2}\|\Delta\|^2}\left(\frac{1}{t}\sum_{i=1}^t \omega_{i,\Delta}^2\right) \geq \frac{D_{\hat{K}}(\Delta)^2}{D_K(\Delta)^2} \geq \left(\frac{1}{t}\sum_{i=1}^t \omega_{i,\Delta}^2\right) - \frac{\|\Delta\|^2}{12}\cdot\frac{1}{t}\sum_{i=1}^t \omega_{i,\Delta}^4. \tag{5}$$

Roadmap. To understand the detailed relative error in $D_K(\Delta)$, what remains is the case when $\|\Delta\|$ is small. As we will start to observe above, when $\|\Delta\|$ is small, then $D_K(\Delta)$ behaves like $\|\Delta\|$ and we can borrow insights from ℓ_2 embeddings. Then combining the two cases (when $\|\Delta\|$ is large and when $\|\Delta\|$ is small) we can achieve "for all bounds" either via simple union bounds, or through a special "continuous" form of net arguments when X is in a bounded range. Similarly, we will show near-matching lower bounds via appealing to near-ℓ_2 properties or via net arguments.

3. Lower Bounds and Relation to ℓ_2 on Small Distances

In this section we show that in the limit as the region containing X shrinks, then all distances act like ℓ_2. This approach is enough for a lower bound, but does not contain the full case $\|\Delta\| \leq 1$, so is not enough for upper bounds.

Lemma 2 *For scalar scaling parameter λ, $\lim_{\lambda \to 0} \frac{D_{\hat{K}}(\lambda\Delta)^2}{D_K(\lambda\Delta)^2} = \frac{1}{t}\sum_{i=1}^{t} \omega_{i,\Delta}^2$.*

Proof Observe that $\omega_{i,\Delta} = \omega_{i,\lambda\Delta}$, for any $\lambda > 0$. Thus in equation (5), $\lim_{\lambda\to 0} 1/(1 - \frac{1}{2}\|\lambda\Delta\|^2)$ goes to 1 so the left hand-side approaches $\frac{1}{t}\sum_{i=1}^{t} \omega_{i,\Delta}^2$. Similarly, $\lim_{\lambda\to 0} \|\lambda\Delta\|^2/12$ goes to 0, and the right-hand side also approaches $\frac{1}{t}\sum_{i=1}^{t} \omega_{i,\Delta}^2$. ■

If we fix Δ then $\omega_{i,\Delta}, 1 \leq i \leq t$ are i.i.d Gaussian variables with mean 0 and standard deviation 1. Thus $\sum_{i=1}^{t} \omega_{i,\Delta}^2$ is a χ^2-variable with t degrees of freedom.

This implies that when $\|x - y\|$ is small, $D_{\hat{K}}(x, y)$ behaves like a Johnson-Lindenstrauss (JL) random projection of $\|x - y\|$, and we can invoke known JL lower bounds.

In particular, Lemma 2 implies if the input data set $X \subset \mathbb{R}^d$ is in a sufficiently small neighborhood of zero, the relative error is preserved only when $\sum_{i=1}^{t} \omega_{i,x,y}^2 \|\lambda(x - y)\|^2 \in [(1 - \epsilon)\|\lambda(x - y)\|^2, (1 + \epsilon)\|\lambda(x - y)\|^2]$ for all $x, y \in X$, and for all arbitrary $\lambda \in \mathbb{R}$. Which implies for arbitrary $x, y \in X$, and $\lambda \in \mathbb{R}$ that

$$\sqrt{\sum_{i=1}^{t} |\omega_i \cdot \lambda(x - y)|^2} = \sqrt{\sum_{i=1}^{t} \omega_{i,x,y}^2 \lambda \|x - y\|^2} \in [(1 - \varepsilon)\,\lambda\,\|x - y\|\,, (1 + \varepsilon)\,\lambda\,\|x - y\|]\,.$$

The far left hand side is in fact the norm $\|g(x) - g(y)\|$ where $g(x)$ is the vector with coordinates $(\omega_1 \cdot \lambda x, ..., \omega_t \cdot \lambda x)$. Thus these are the exact conditions for relative error bounds on embedding ℓ_2 via the Johnson-Lindenstrauss transforms, which gives the following.

Lemma 3 *If for any $n, d > 0, X \subset \mathbb{R}^d$ s.t. $|X| = n$, using $t(n, \varepsilon)$ pairs of random Fourier features, $\frac{D_{\hat{K}}(x,y)}{D_K(x,y)} \in [1 - \varepsilon, 1 + \varepsilon]$ with probability $1 - \delta$, then there exists a random linear embedding with $t(n, \varepsilon)$ projected dimensions preserving the ℓ_2-norm for all pairs $x, y \in S$ up to relative error with probability at least $1 - \delta$.*

Theorem 4 *There exists a set of n points $X \subset \mathbb{R}^d$ so that $t = \Omega(\frac{1}{\varepsilon^2}\log n)$ pairs of random features (hence $m = 2t$ dimensions), for any $\varepsilon \in (0, 1/2)$, are necessary so for any $x, y \in X$ that $\frac{D_{\hat{K}}(x,y)}{D_K(x,y)} \in [1 - \varepsilon, 1 + \varepsilon]$.*

Proof A lower bound of $\Omega(\frac{1}{\varepsilon^2}\log n)$ projected dimensions for linear embeddings in ℓ_2 is here Larsen and Nelson (2017). ■

4. Relative Error Bounds For Small Distances and Small Data Sets

The Taylor expansion in equation (5) and additive errors via equation (2) are only sufficient to provide us bounds for $\|\Delta\| \leq O(\sqrt{\varepsilon}/\log(1/\varepsilon))$ or for $\|\Delta\| \geq 1$. In this section we need to use a more powerful technique or moment generating functions to fill in this gap.

In particular, $1-\cos(\omega_{i,\Delta}\|\Delta\|)$ is a sub-Gaussian random variable so it is expected to have a good concentration, but it is not enough to follow the idea crudely if we want relative error bounds; we derive a more precise bound of the moment generating function of $1-\cos(\omega_{i,\Delta}\|\Delta\|)$ as follows.

Lemma 5 *For $\omega \sim \mathcal{N}(0,1)$ and $0 \leq \|\Delta\| \leq 1$, let $M(s)$ be the moment generating function of $1-\cos(\omega\|\Delta\|) - \left(1-e^{-\frac{1}{2}\|\Delta\|^2}\right) = e^{-\frac{1}{2}\|\Delta\|^2} - \cos(\omega\|\Delta\|)$. Then for all $s \geq 0$, $\ln M(s) \leq \frac{1}{4}s^2\|\Delta\|^4$.*

This technical proof is in Appendix A. We next combine this result with an existing bound on sub-exponential random variables Buldygin et al. (2000)[Lemma 4.1 in Chapter 1]. Let X be a random variable, and let $M(s)$ be the moment generating function of $X - \mathbf{E}[X]$. Let $\bar{X}_t := \frac{1}{t}\sum_{i=1}^{t} X_i$ where $X_1, ..., X_t$ are i.i.d. samples of X. If $\ln M(s) \leq \frac{s^2 p}{2}$ for all $s \in [0, \frac{1}{q})$, then

$$P(|\bar{X}_t - \mathbf{E}[X]| \geq \varepsilon \mathbf{E}[X]) \leq 2\exp\left(-\min\left(t\frac{\varepsilon^2 \mathbf{E}[X]^2}{2p}, t\frac{\varepsilon \mathbf{E}[X]}{2q}\right)\right). \tag{6}$$

Lemma 6 *If $\|x-y\| \leq \sigma$, $t = \Omega(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$, then $\mathbf{Pr}\left(\frac{D_{\hat{K}}(x,y)}{D_K(x,y)} \in [1-\varepsilon, 1+\varepsilon]\right) \geq 1-\delta$.*

Proof Recall that $\langle \hat{f}_i(x), \hat{f}_i(y)\rangle = \cos(\langle \omega_i, (x-y)\rangle)$ and $(1/2)D_{\hat{K}}(x,y)^2 = \frac{1}{t}\sum_{i=1}^{t}(1-\cos(\langle \omega_i(x-y)\rangle))$. Then define random variable $X_i = (1-\cos(\langle \omega_i(x-y)\rangle))$, and $X = \frac{1}{t}\sum_{i=1}^{t} X_i$. Since $\mathbf{E}[X] = \mathbf{E}[D_{\hat{K}}(x,y)^2] = D_K(x,y)^2$, then $\mathbf{E}[X_i] = 1-\exp(-\frac{1}{2}\|\Delta\|^2)$.

For $M(s)$ the moment generating function of $X_i - \mathbf{E}[X_i]$, by Lemma 5 we have $\ln(M(s)) \leq \frac{1}{2}s^2 p$ for $p = \frac{1}{2}\|\Delta\|^4$ for $s \in [0, \frac{1}{q}]$ with $q = 2\|\Delta\|^2$. Also recall by equation (3) we have for any $x, y \in \mathbb{R}^d$ with $\|x-y\| \leq \sigma$, that $0.86 \leq \frac{D_K(x,y)}{\|\Delta\|} \leq 1$.

Plugging these values into equation (6) with $t = \frac{6}{\varepsilon^2}\frac{\|\Delta\|^4}{D_K(x,y)^4}\ln(2/\delta) = O(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$, we obtain that

$$\begin{aligned}
\mathbf{Pr}[|D_{\hat{K}}(x,y) - D_K(x,y)| \geq \varepsilon D_K(x,y)] &= \mathbf{Pr}[|X - \mathbf{E}[X]| \geq \varepsilon \mathbf{E}[X]] \\
&\leq 2\exp\left(-\min\left(t\frac{\varepsilon^2 \mathbf{E}[X]^2}{\|\Delta\|^4}, t\frac{\varepsilon \mathbf{E}[X]}{4\|\Delta\|^2}\right)\right). \\
&= 2\exp\left(-\min\left(6, \frac{3}{2\varepsilon}\frac{\|\Delta\|^2}{D_K(x,y)^2}\right)\ln\frac{2}{\delta}\right) \\
&\leq 2\exp\left(-\min\left(6, \frac{1}{\varepsilon}\right)\ln\frac{2}{\delta}\right) \leq \delta.
\end{aligned}$$

■

Together with Lemma 1 (for $\|x-y\| \geq \sigma$), we apply a union bound over all $\binom{n}{2}$ pairs vectors from a set of n vectors.

Theorem 7 *For any set $X \subset \mathbb{R}^d$ of size n, then $m = 2t = \Omega(\frac{1}{\varepsilon^2}\log n)$ projected dimensions are sufficient so $\frac{D_{\hat{K}}(x,y)}{D_K(x,y)} \in [1-\varepsilon, 1+\varepsilon]$ with high probability (e.g., at least $1-1/n$).*

5. Relative Error Bounds for Low Dimensions and Diameter

Here we prove that the relative error bound holds for the infinitely many pairs of vectors of finite distance to each other, given that the number of dimensions is small. A common approach in subspace embeddings replaces n with the size of a sufficiently fine net; given a smoothness condition, once the error is bounded on the net points, the guarantee is extended to the 'gaps' in between.

On the other hand, the Gaussian kernel distance is non-linear, so it is not immediately clear how the above technique can apply. We begin with the Lipschitz constant of $D_{\hat{K}}(\cdot)^2$, with respect to the vector Δ, *not individual points in* $\mathbb{R}^d$. Then we develop a fine-grained structure and a net on the set of directions $\Delta/\|\Delta\|$ as long as $\|\Delta\|$ is small enough, using an object we call a λ-urchin.

Lipschitz bound. First we provided the needed Lipschitz bound with respect to Δ.

Lemma 8 *For any $\Delta \in \mathbb{R}^d$, $|\nabla D_{\hat{K}}(\Delta)^2| \le 2\frac{1}{t}\sum_{i=1}^{t} \|\omega_i\|_1 \|\omega_i\| \|\Delta\|$.*

Proof We denote by $\omega_i^{(j)}$ the jth coordinate of ω_i; where recall $\omega_{i,\Delta} = \langle \omega_i, \Delta \rangle$.

$$\left|\nabla D_{\hat{K}}(\Delta)^2\right| = 2\left|\frac{1}{t}\sum_{i=1}^{t}\sum_{j=1}^{d}\omega_i^{(j)}\sin(\langle\omega_i,\Delta\rangle)\right| \le 2\frac{1}{t}\sum_{i=1}^{t}\sum_{j=1}^{d}|\omega_i^{(j)}|\,|\sin(\langle\omega_i,\Delta\rangle)|$$
$$\le 2\frac{1}{t}\sum_{i=1}^{t}\|\omega_{i,\Delta}\|_1|\langle\omega_i,\Delta\rangle| \le 2\frac{1}{t}\sum_{i=1}^{t}\|\omega_i\|_1\|\omega_i\|\|\Delta\| \qquad \blacksquare$$

Corollary 9 *For any $c \ge 0$, over the region $\|\Delta\| \le c$, the Lipschitz constant of $D_{\hat{K}}(\Delta)^2$ is bounded above by $O(c \cdot \sqrt{d}\log(d/\delta))$ with probability at least $1 - O(\delta)$.*

Proof We can bound any coordinate $\omega_i^{(j)}$ of ω_i so that $|\omega_i^{(j)}| \le O(\log\frac{1}{\delta})$ with probability at least $1-\delta$. By a union, bound the all coordinates $|\omega_i^{(j)}| \le O(\log\frac{d}{\delta})$ with probability at least $1-\delta$. So the gradient is bounded by $2\|\Delta\|\frac{1}{t}\sum_{i=1}^{t}\|\omega_i\|_1\|\omega_i\| \le \|\Delta\|\sqrt{d}O(\log\frac{1}{\delta}) \le O(c\cdot\sqrt{d}\log\frac{1}{\delta})$, which also bounds the Lipschitz constant. $\blacksquare$

In case $c = \frac{\sqrt{\varepsilon}}{2\ln(4/\varepsilon\delta)}$, the Lipschitz constant is $O(\sqrt{\varepsilon d}\log\frac{d}{\delta})$ with probability at least $1-\delta$.

Fine-grained small distance structure. We now analyze equation (5). We first state a standard bound on χ^2 random variables $\sum_{i=1}^{t}\omega_{i,\Delta}^2$, and then show how to bound the other terms.

Lemma 10 *For $\varepsilon \in (0,1)$, $\delta \in (0,\frac{1}{2})$, if $t \ge 8\frac{1}{\varepsilon^2}\ln(2/\delta)$ then* $\mathbf{Pr}\left[\frac{1}{t}\sum_{i=1}^{t}\omega_{i,\Delta}^2 \notin [1-\varepsilon, 1+\varepsilon]\right] \le \delta$.

Proof Here we use Lemma 1 from B. Laurent (2000); if X is a χ^2 random variable with t degrees of freedom $\mathbf{Pr}[t-2\sqrt{tx} \le X \le t+2\sqrt{tx}+2x] \ge 1-2e^{-x}$. We can set $x = \frac{1}{8}t\varepsilon^2$ then $t-2\sqrt{tx} = t-\varepsilon t/\sqrt{2}$, and $t+2\sqrt{tx}+2x = t+\varepsilon t/\sqrt{2}+\frac{1}{4}\varepsilon^2 t < t+\varepsilon t$. Also, $2e^{-x} = 2e^{-\frac{1}{8}t\varepsilon^2} = 2e^{-\ln(2/\delta)} = \delta/2 \le \delta$ for $\delta \le 1/2$. Therefore, $\frac{1}{t}\sum_{i=1}^{t}\omega_{i,\Delta}^2 \notin [1-\varepsilon,1+\varepsilon]$ with probability at most δ. $\blacksquare$

Now to bound the other parts ($\|\Delta\|^2/2$ and the term containing $\omega_{i,\Delta}^4$) of equation (5) requires a further restriction on $\|\Delta\|$.

Lemma 11 *For $\varepsilon \in (0,1)$ and $\delta \in (0,2/5)$, if $\|\Delta\| \le \frac{\sqrt{\varepsilon}}{2\ln(4/\varepsilon\delta)}$ for a constant C, and $t \ge 18\frac{1}{\varepsilon^2}\ln(4/\delta)$, then with probability at least $1-\delta$, for all $\lambda \in [0,1]$ we have $\frac{D_{\hat{K}}(\lambda\cdot\Delta)^2}{D_K(\lambda\cdot\Delta)^2} \in [1-\varepsilon, 1+\varepsilon]$.*

Proof If ω is a standard Gaussian variable, then $|\omega| \le \sqrt{2\ln(1/\delta')}$ with probability at least $1-\delta'$. Using $\delta' = \delta/2t$, then applying a union bound ensures that (using $\ln(4/\delta) < 1/\delta$ for $\delta < 2/5$)

$$\omega_{i,\Delta} \le \sqrt{2\ln(2t/\delta)} = \sqrt{2\ln(\frac{16}{\delta\varepsilon^2}\ln(4/\delta))} \le \sqrt{2\ln(\frac{16}{\delta^2\varepsilon^2})} = 2\sqrt{\ln(4/\delta\varepsilon)},$$

for t such random variables with probability at least $1-(\delta' t) = 1-\delta/2$. This means, if $\|\Delta\| \le \frac{\sqrt{\varepsilon}}{2\ln(4/\varepsilon\delta)}$ then $\omega_{i,\Delta}\|\Delta\| \le \sqrt{\frac{\varepsilon}{\ln(4/\varepsilon\delta)}}$ with probability at least $1-\delta/2$ for t such random variables. Also then each $\omega_{i,\Delta}\|\Delta\| \le 1$ satisfies the conditions for (5).

Then using $\|\Delta\| \le \frac{\sqrt{\varepsilon}}{\log(1/\varepsilon\delta)}$ and $\omega_{i,\Delta} \le 2\sqrt{\ln(4/\varepsilon\delta)}$ we can bound the last term in (5) as

$$\frac{\|\Delta\|^2}{12}\cdot\frac{1}{t}\sum_{i=1}^{t}\omega_{i,\Delta}^4 = \frac{\|\Delta\|^2}{12}\cdot\left(2\sqrt{\ln(4/\varepsilon\delta)}\right)^4 \le \frac{\varepsilon}{12\cdot 4\ln^2(4/\varepsilon\delta)}\cdot 16\ln^2(4/\varepsilon\delta) = \frac{\varepsilon}{3}.$$

Then along with Lemma 10 (error $1-\frac{2\varepsilon}{3}$) and RHS of (5) (error $\frac{\varepsilon}{3}$), we have $\frac{D_{\hat{K}}(\Delta)^2}{D_K(\Delta)^2} \ge 1-\varepsilon$.

Similarly, Lemma 10 and $\frac{1}{2}\|\Delta\|^2 < \frac{\varepsilon}{8\ln^2(4/\varepsilon\delta)} < \frac{\varepsilon}{8}$ imply the LHS of (5) is bounded above by $(1+\frac{2}{3}\varepsilon)(1/(1-\frac{\varepsilon}{8})) \le 1+\varepsilon$ with probability at least $1-\delta/2$. Thus, we have $\frac{D_{\hat{K}}(\Delta)^2}{D_K(\Delta)^2} \in [1-\varepsilon, 1+\varepsilon]$.

For $\frac{D_{\hat{K}}(\lambda\Delta)^2}{D_K(\lambda\Delta)^2} \in [1-\varepsilon,1+\varepsilon]$, note that the above analysis still holds if we scale $\|\Delta\|$ to be smaller, i.e. as long as $\lambda \in [0,1]$. In particular, $\omega_{i,\Delta}$ is unchanged by the scaling λ. ■

Scaled net argument. We can now provide a net argument for a relative error bound for all small Δ. Intuitively, what separates typical net arguments from ours is the scaling λ in Lemma 11; our 'net' contains a set of line segments extending from the origin, which we call a *λ-urchin*.

Lemma 12 *If $t = \Omega(\frac{d}{\varepsilon^2}\log\left(\frac{d}{\varepsilon}\frac{1}{\delta}\right))$, then with probability at least $1-\delta$, for all Δ such that $\|\Delta\| \le \frac{\sqrt{\varepsilon}}{2\ln(4/\varepsilon\delta)}$, then $\frac{D_{\hat{K}}(\Delta)^2}{D_K(\Delta)^2} \in [1-\varepsilon,1+\varepsilon]$.*

Proof The proof will first consider distances Δ such that $\{\Delta : \|\Delta\| = R_\varepsilon\}$ where $R_\varepsilon = \frac{\sqrt{\varepsilon}}{2\ln(4/\varepsilon\delta)}$, and then generalize to smaller distances using Lemma 11 and a construction we call a λ-urchin.
Fixed distance case: Consider two points Δ_1, Δ_2 from the surface $\{\Delta : \|\Delta\| = R_\varepsilon\}$. If $\|\Delta_1 - \Delta_2\| \le \frac{\sqrt{\varepsilon}}{\sqrt{d}\log\frac{1}{\delta}}R_\varepsilon^2$ then Corollary 9 implies

$$\begin{aligned}\left|D_{\hat{K}}(\Delta_1)^2 - D_{\hat{K}}(\Delta_2)^2\right| &\le O(\sqrt{\varepsilon d}\log\frac{1}{\delta})\cdot\|\Delta_1-\Delta_2\| \le O(\sqrt{\varepsilon d}\log\frac{1}{\delta})\cdot\frac{\sqrt{\varepsilon}}{\sqrt{d}\log\frac{1}{\delta}}R_\varepsilon^2 \\ &= O(\varepsilon\cdot R_\varepsilon^2) = O(\varepsilon)\cdot D_K(\Delta_1)^2.\end{aligned}$$

Now let Γ_γ be a γ-net over $\{\Delta : \|\Delta\| = R_\varepsilon\}$ where $\gamma \le \frac{\sqrt{\varepsilon}}{\sqrt{d}\log\frac{d}{\delta}}R_\varepsilon^2$. For any $\Delta_1 \in \{\Delta : \|\Delta\| = R_\varepsilon\}$, there exists $\Delta_2 \in \Gamma_\gamma$ such that $\|\Delta_1 - \Delta_2\| \le \gamma$. Then the above implies

$$(1-O(\varepsilon))D_{\hat{K}}(\Delta_2)^2 \le D_{\hat{K}}(\Delta_1)^2 \le (1+O(\varepsilon))D_{\hat{K}}(\Delta_2)^2. \tag{7}$$

By the triangle inequality, equation (3), and $\sqrt{d}\log\frac{d}{\delta}\cdot 2\ln(4/\varepsilon\delta) > 1$, we have

$$\begin{aligned}|D_K(\Delta_1) - D_K(\Delta_2)| &\le D_K(\Delta_1,\Delta_2) \le \|\Delta_1-\Delta_2\| \le \gamma \\ &\le \frac{\sqrt{\varepsilon}}{\sqrt{d}\log\frac{d}{\delta}}R_\varepsilon^2 = \varepsilon\frac{1}{\sqrt{d}\log\frac{d}{\delta}\cdot 2\ln(4/\varepsilon\delta)}R_\varepsilon < \varepsilon\cdot O(D_K(\Delta_1)).\end{aligned} \tag{8}$$

We will choose $t = \Omega(\frac{1}{\varepsilon^2}\log\frac{|\Gamma_\gamma|}{\delta})$ so the following holds over Γ_γ with probability at least $1-\delta$

$$(1-O(\varepsilon))D_K(\Delta_2)^2 \le D_{\hat{K}}(\Delta_2)^2 \le (1+O(\varepsilon))D_K(\Delta_2)^2. \tag{9}$$

These equations (7), (9), and (8) show, respectively that the ratios $\frac{D_{\hat{K}}(\Delta_1)}{D_{\hat{K}}(\Delta_2)}$, $\frac{D_{\hat{K}}(\Delta_2)}{D_K(\Delta_2)}$, and $\frac{D_K(\Delta_2)}{D_K(\Delta_1)}$ are all in $[1+O(\varepsilon), 1-O(\varepsilon)]$; hence we can conclude

$$|D_K(\Delta_1) - D_{\hat{K}}(\Delta_1)| \le O(\varepsilon)\cdot D_K(\Delta_1). \tag{10}$$

Which are in turn $1 \pm O(\varepsilon)$ relative error bounds for the kernel distance, over $\{\Delta : \|\Delta\| = R_\varepsilon\}$.

All distances case: For the region $\{\Delta : \|\Delta\| < R_\varepsilon\}$, consider again Γ_γ. For each net point $p \in \Gamma_\gamma$ we draw a line segment from p to the origin, producing the set of line segments $\bar{\Gamma}_\gamma$, that we call the γ-*urchin*. By Lemma 11, and $t = \Omega(\frac{1}{\varepsilon^2}\log\frac{|\Gamma_\gamma|}{\delta})$, with probability at least $1-\delta$, we have relative error bounds for the Gaussian kernel distance over the γ-urchin.

Now for any $\lambda \in (0,1)$, consider the intersection $\{\Delta : \|\Delta\| = \lambda R_\varepsilon\} \cap \bar{\Gamma}_\gamma$. We see that the γ-urchin induces a net over $\{\Delta : \|\Delta\| = \lambda R_\varepsilon\}$. Due to scaling we can see that, in fact, it is a $(\lambda\gamma)$-net. So the distance between any point in $\{\Delta : \|\Delta\| = \lambda R_\varepsilon\}$ and the closest net point is bounded above by $\frac{\lambda\sqrt{\varepsilon}}{\sqrt{d}\log\frac{d}{\delta}}R_\varepsilon^2$. From Corollary 9, the Lipschitz constant is now $O(\lambda\cdot\sqrt{\varepsilon d})$.

By arguments similar to those leading to (10) we obtain, for any $\Delta_1 \in \{\Delta : \|\Delta\| = \lambda R_\varepsilon\}$

$$|D_K(\Delta_1) - D_{\hat{K}}(\Delta_1)| \le O(\varepsilon)\cdot\lambda\cdot R_\varepsilon \le O(\varepsilon)\cdot D_K(\Delta_1). \tag{11}$$

Since this holds for all $\lambda \in [0,1]$, we obtain relative error bounds over $\{\Delta : \|\Delta\| \le R_\varepsilon\}$.

The size of Γ_γ is bounded above by $O((R_\varepsilon\frac{1}{\gamma})^d) = O((R_\varepsilon\cdot\frac{\sqrt{d}\log\frac{d}{\delta}}{\sqrt{\varepsilon}}\frac{1}{R_\varepsilon^2})^d) = O((\frac{\sqrt{d}\log\frac{d}{\delta}}{\sqrt{\varepsilon}}\frac{1}{R_\varepsilon})^d) = O((\frac{\sqrt{d}\log\frac{d}{\delta}\log\frac{1}{\varepsilon\delta}}{\varepsilon})^d)$. It is sufficient to have $t = O(\frac{1}{\varepsilon^2}\log\frac{|\Gamma_\gamma|}{\delta}) = O(\frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta})$ so that relative error holds over the γ-net and the γ-urchin simultaneously, which imply (11) and (10), with probability at least $1-\delta$. ■

Corollary 13 *If $t = \Omega(\frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta})$, then for all Δ such that $\|\Delta\| \le 1$, $\frac{D_{\hat{K}}(\Delta)}{D_K(\Delta)} \in [1-\varepsilon, 1+\varepsilon]$ with probability at least $1-\delta$.*

Proof Consider the region $1 \ge \|\Delta\| > \frac{\sqrt{\varepsilon}}{2\ln(4/\varepsilon\delta)}$. The Lipschitz constant is bounded above by $O(t\sqrt{d}\log\frac{d}{\delta})$ by Corollary 9, so we only need a γ-net where $\gamma \le \frac{\varepsilon^2}{t\sqrt{d}\log\frac{d}{\delta}}$ to give relative error by standard net arguments. The size of this net is at most $\left(\frac{\sqrt{d}\log(d/\delta)}{\varepsilon^2}\right)^d$, so again it suffices to set $t = O(\frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta})$ for our embeddings as above. ■

Combined with Lemma 1 for $\|\Delta\| > 1$ we obtain:

Theorem 14 *If $t = \Omega\left(\frac{d}{\varepsilon^2}\log\left(\frac{d}{\varepsilon}\frac{\mathcal{M}}{\delta}\right)\right)$, then for any $\mathcal{M} \ge 0$, $\frac{D_{\hat{K}}(x,y)^2}{D_K(x,y)^2} \in [1-\varepsilon, 1+\varepsilon]$ holds for all $x, y \in \mathbb{R}^d$ such that $\|x-y\|/\sigma \le \mathcal{M}$ with probability at least $1-\delta$.*

Proof Set $t = \Omega(\frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta}) + \Omega(\frac{d}{\varepsilon^2}d\log\frac{d\mathcal{M}}{\varepsilon\delta}) = \Omega\left(\frac{d}{\varepsilon^2}\log\left(\frac{d}{\varepsilon}\frac{\mathcal{M}}{\delta}\right)\right)$ to account for both cases $\|\Delta\| = \frac{\|x-y\|}{\sigma} \le 1$ and $1 \le \frac{\|x-y\|}{\sigma} \le \mathcal{M}$, respectively. ■

6. Lower Bounds for Low Dimensions

When is n is unbounded, a recent paper Sriperumbudur and Szabo (2015) implies that, even for small d, $D_{\hat{K}}$ cannot $(1+\varepsilon)$-approximate D_K unless $\mathcal{M}$ is bounded. Here we provide an explicit and *general* lower bound depending on $\mathcal{M}$ and d that matches the our upper bound up to a $O(\log \frac{1}{\varepsilon})$ factor.

First we need the following general result Alon (2003)[Theorem 9.3] related to embedding to ℓ_2. Let B be an $n \times n$ real matrix with $b_{i,i} = 1$ for all i and $|b_{i,j}| \leq \varepsilon$ for all $i \neq j$. If the rank of B is r, and $\frac{1}{\sqrt{n}} < \varepsilon < 1/2$, then $r \geq \Omega(\frac{1}{\varepsilon^2 \log(1/\varepsilon)} \log n)$. Geometrically, r is the minimum number of dimensions that can contain a set of n near-orthogonal vectors. Indeed, any set S of n near-orthogonal vectors can be rotated to form the rows of a matrix of the form of B, and the rank is then the lowest number of dimensions that contain S.

Lemma 15 *Given $\mathcal{M} \geq 0$, let $B_{\mathcal{M}}(0)$ be the ball in $\mathbb{R}^d$ centered at the origin with radius $\mathcal{M}$. Let $h : \mathbb{R}^d \to \mathbb{R}^t$ be a mapping such that for any $x \neq y \in B_{\mathcal{M}}(0)$ we have $|K(x,y) - h(x) \cdot h(y)| \leq \varepsilon \leq \frac{1}{4}$. Then with sufficiently large $\mathcal{M}$, $t = \Omega(\frac{d}{\varepsilon^2 \log(1/\varepsilon)} \log(\frac{\mathcal{M}}{\log(1/\varepsilon)}))$.*

Proof Consider a subset $S \subset \mathbb{R}^d$ in $B_{\mathcal{M}}(0)$ so for all $x, y \in S$, with $x \neq y$, we have $\|x - y\| \geq \sigma\sqrt{2 \log \frac{1}{\varepsilon}}$. Then for any $x, y \in S$, $K(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2}) \leq \varepsilon$. In particular, define S as the intersection of $B_{\mathcal{M}}(0)$ with an orthogonal grid of side length $\sigma\sqrt{2\log(1/\varepsilon)}$; it has size $\Omega\left(\left(\frac{\mathcal{M}}{\log(1/\varepsilon)}\right)^d\right)$.

For any $x, y \in S$, $|h(x) \cdot h(y)| \leq 2\varepsilon$, and also $|\{h(s) \mid s \in S\}| = |S|$. Then Alon (2003)[Theorem 9.3] implies the dimension of h must be $t = \Omega(\frac{1}{\varepsilon^2 \log(1/\varepsilon)} \log |S|) = \Omega(\frac{d}{\varepsilon^2 \log(1/\varepsilon)} \log(\frac{\mathcal{M}}{\log(1/\varepsilon)}))$. ■

Theorem 16 *Given $\mathcal{M} \geq 0$, let $B_{\mathcal{M}}(0)$ be the ball in $\mathbb{R}^d$ centered at the origin with radius $\mathcal{M}$. Let $h : \mathbb{R}^d \to \mathbb{R}^t$ be a mapping such that for any $x, y \in B_{\mathcal{M}}(0)$ we have $1 - \varepsilon \leq \frac{D_K(x,y)}{\|h(x)-h(y)\|} \leq 1 + \varepsilon$ with $\varepsilon \leq \frac{1}{4}$. Restrict that for any $x \in \mathbb{R}^d$ that $\|h(x)\| = 1$. If $\mathcal{M}$ is sufficiently large, $t = \Omega(\frac{d}{\varepsilon^2 \log(1/\varepsilon)} \log(\frac{\mathcal{M}}{\log(1/\varepsilon)}))$.*

Proof Consider a set (as in proof of Lemma 15) $S \subset B_{\mathcal{M}}(0)$. If for all $x, y \in S$ we have $1 - \varepsilon \leq \frac{D_K(x,y)}{\|h(x)-h(y)\|} \leq 1 + \varepsilon$, then it implies

$$|D_K(x,y)^2 - \|h(x) - h(y)\|^2| \leq \Theta(\varepsilon) D_K(x,y)^2 \leq \Theta(\varepsilon),$$

since $D_K(x,y) < 2$. Expanding $D_K(x,y)^2 = 2 - 2K(x,y)$ and $\|h(x) - h(y)\|^2 = 2 - 2\langle h(x), h(y)\rangle$ implies that $|K(x,y) - \langle h(x), h(y)\rangle| \leq \Theta(\varepsilon)$ as well. However, Lemma 15 implies that for sufficiently small ε (adjusting the constant in $\Theta(\varepsilon)$) that we require the $t = \Omega(\frac{d}{\varepsilon^2 \log(1/\varepsilon)} \log(\frac{\mathcal{M}}{\log(1/\varepsilon)}))$. ■

This implies the impossibility of fully embedding into ℓ_2 the Gaussian kernel distance over the *entire* $\mathbb{R}^d$, i.e. for an infinite number of points, answering a question raised by Sriperumbudur and Szabo (2015). This argument can also extend to show a dependency on $d \log \mathcal{M}$ is inevitable when we do not have a bound on n.

7. Discussion

We have demonstrated theoretically tight relative error for kernel distance using random Fourier features, indicating tighter approximations for several important learning applications. In the following, we make some further remarks on the implications of our results, and then also empirically observe these properties.

7.1. Implications in Learning and Analysis

In addition to the applications of our bounds to oblivious kernel embeddings, kernel k-means clustering, and kernel distance matching that we discussed in Section 1.3, we mention a couple more below.

Geometric approximation in learning and mining. Our results show that random feature mappings allow for a finer notion of approximating the geometry of RKHS than previously known. In particular, our low-dimensional bounds in Section 5, imply that if an object $U \subset \mathbb{R}^d$ (such as a non-linear decision boundary) and training data $S \subset \mathbb{R}^d$ both lie within a ball with finite radius $\mathcal{M}$, then for any point $x \in S$, the minimum kernel distance between U and x is approximately preserved in the random feature space as $\min_{y \in U} \|\phi(x) - \phi(y)\|$. For instance "large-margin" techniques and analyses Tsochantaridis et al. (2005) condition on the margin $\gamma = \max_{x \in S} \min_{y \in U} \|\phi(x) - \phi(y)\|$ being large, so we also preserve relative errors on this margin. This suggests better performance guarantees for kernelized learning large-margin techniques, and those involving the minimization of ℓ_2 distances, such as in kernel SVM (hinge-loss) and in kernel PCA (recovery error); see Section 8.

Gram matrix approximation. The approximation error of inner products is proportional to the approximation error of distances. This is because both ϕ and $\hat{\phi}$ map every input point to a unit vector; thus $D_K(x,y)^2 = 2 - 2K(x,y)$ and $D_{\hat{K}}(x,y)^2 = 2 - 2\hat{K}(x,y)$, for any distinct $x, y \in \mathbb{R}^d$. Therefore $|K(x,y) - \hat{K}(x,y)|$ is the same as $\frac{1}{2}|D_K(x,y)^2 - D_{\hat{K}}(x,y)^2|$. Hence approximation error of the Gram matrix is bounded in terms of the sum of pairwise squared distances

$$\|G - \hat{G}\|_1 \leq \frac{1}{2} \sum_{x \in X} \sum_{y \in X} |D_K(x,y)^2 - D_{\hat{K}}(x,y)^2| \leq \varepsilon \sum_{x \in X} \sum_{y \in X} D_K(x,y)^2,$$

with high probability, when m is set for the appropriate data setting in our bounds. Thus we have in some sense sharper bounds on approximating the Gram matrix.

7.2. Remark on Lower Bound in n

A new result of Larsen and Nelson (2017) provides a $t = \Omega(\frac{1}{\varepsilon^2} \log n)$ lower bound for even *non-linear* embeddings of a size n point set in $\mathbb{R}^d$ into $\mathbb{R}^t$ that preserve distances within $(1 \pm \varepsilon)$. It holds for any $\varepsilon \in (1/\min\{n,d\}^{0.4999}, 1)$. Since, there exists an isometric embedding of any set of n points in any RKHS into $\mathbb{R}^n$, then this $t = \Omega(\frac{1}{\varepsilon^2} \log n)$ lower bound suggests that it applies to $\hat{\phi}$ and $\tilde{\phi}$ or any other technique, for almost any ε. However, it is not clear that *any* point set (including the ones used in the strong lower bound proof Larsen and Nelson (2017)), can result from an isomorphic (or approximate) embedding of RKHS into $\mathbb{R}^n$. Hence, this new result does not immediately imply the lower bound we show in Section 3.

Moreover, the proof of Theorem 4 retains two points of potential interest. First it holds for a (very slightly) larger range of $\varepsilon \in (0,1)$. Second, Lemma 3 highlights that at very small ranges, $\hat{\phi}$ is indistinguishable from the standard JL embedding.

8. Empirical Demonstration of Relative Error

We demonstrate that relative error actually results from the $\hat{\phi}$ kernel embeddings in two ways. First we demonstrate relative error bounds for kernel PCA. Second we show this explicitly for pairwise distances in the embedding.

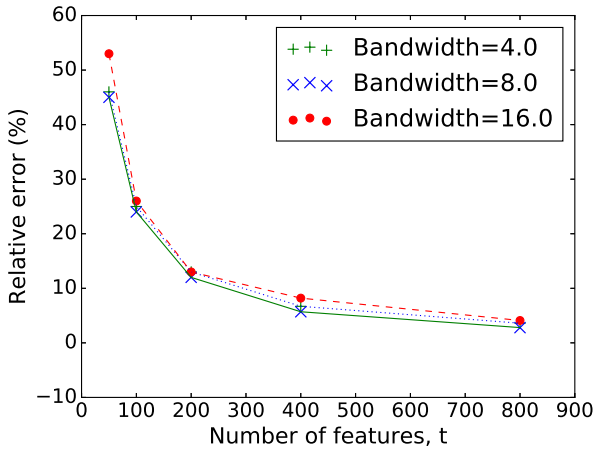
8.1. Relative Error for Kernel PCA

When PCA is applied to approximate a data matrix in practice, the allowed approximation error is often chosen to be a small but constant (e.g. 10%) fraction of the total variance. Our results imply relative error in the approximation of the total variance, so we can also show relative error in typical cases of performing kernel PCA with Gaussian kernels using Random Fourier Features.

We consider two ways of running kernel PCA on the USPS data. By default we use the first $n = 2000$ data points in $\mathbb{R}^d$ for $d = 256$, the first $n/10$ data points of each digit. In the first way, we create the $n \times n$ (centered) gram matrix G of all inner products, and then use the top k eigenvectors to describe the best subspace of RKHS to represent the data; this is treated as a baseline. Second we embed each point into $\mathbb{R}^m$ using $\hat{\phi}$, generating an $n \times m$ matrix Q (after centering). The top k right singular values V_k of Q describe the kernel PCA subspace.

Error in PCA is typically measured as the sum of squared residuals, that is for each point $q \in Q \subset \mathbb{R}^m$, its projection onto V_k is $V_k^T V_k q$, and its residual is $r_q = \|q - V_k^T V_k q\|^2$. Thus r_q is precisely the squared kernel distance between q and its projection. And then the full error is $\hat{R}_k = \|Q - V_k^T V_k Q\|_F^2 = \sum_{q \in Q} \|q - V_k^T V_k q\|^2$. For the non-approximate case, it can be calculated as the sum of eigenvalues in the tail $R_k = \sum_{i=k+1}^{n} \lambda_i$.

Given R_k and $\hat{R}_k$ we can measure the relative error as $\hat{R}_k / R_k$. Our analysis indicates this should be in $[1 - \varepsilon, 1 + \varepsilon]$ using roughly $t = C/\varepsilon^2$ features, where C depends on n or $d \log \mathcal{M}$. To isolate ε we calculate $|\frac{\hat{R}_k}{R_k} - 1|$ averaged over 10 trials in the randomness in $\hat{\phi}$. This is shown in Figure 1 using $k = 40$, with $\sigma \in \{4, 8, 16\}$ and varying $t \in \{50, 100, 200, 400, 800\}$. We observe that our measured error decreases quadratically in t as expected. Moreover, this rate is stable as a function of σ as would be expected where the correct way to quantify error is the relative error we measure.



		$\sigma = 4$	$\sigma = 8$	$\sigma = 16$
Baseline		1667.1	882.5	206.1
t	50	897.6	489.2	96.9
	100	1257.6	666.5	152.2
	200	1453.6	776.4	178.7
	400	1554.7	831.5	189.2
	800	1606.8	857.3	197.6

Figure 1: Relative error $|\frac{\hat{R}k}{R_k} - 1|$ in % , against t, with $n = 2000$, $k = 40$ and different bandwidths. Relative error is roughly stable across different values of σ, and consistently reduced by increasing t.

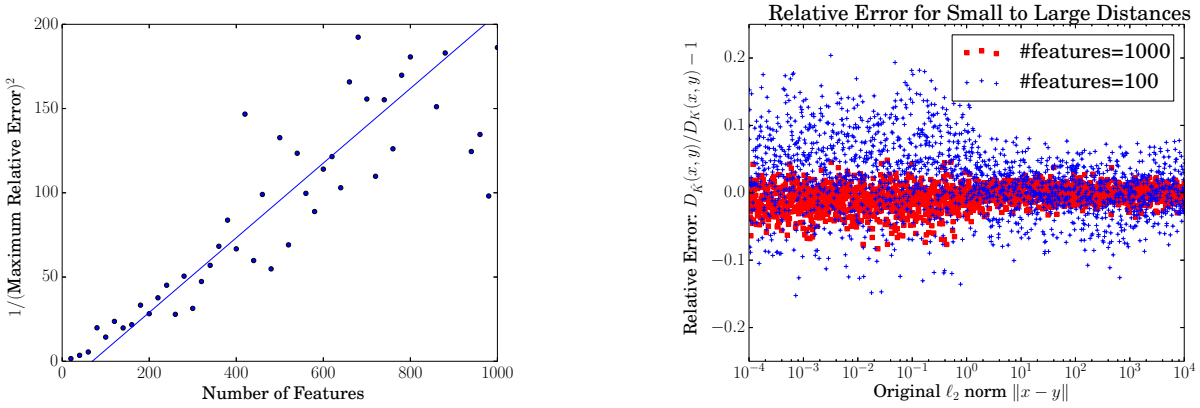


Figure 2: (left) Inverse squared relative errors. (right) Relative errors with varying distance.

8.2. Pairwise Demonstrations of Relative Error

Here we provide simulations that confirm our theoretical findings. We randomly generate pairs of points $(x_1, y_1) \ldots (x_n, y_n)$ with varying ℓ_2 distance $\|x_i - y_i\|$; in particular, x_i is a random point in a ball or radius 500 and y_i is generated to be a random point in the sphere $\|x - y\| = r_i$ where $r_1, ..., r_n$ follow a geometric distribution, ranging from approximately 10^{-4} to 10^4.

In Figure 2(left), for different values of t (the number of features) we generate a fresh sequence of 2000 random pairs, and record the maximum relative error $\varepsilon_{\max} = \max_i \frac{D_K(x_i,y_i)}{\|\phi(x_i)-\phi(y_i)\|}$. The graph shows that t is roughly proportional to $\varepsilon_{\max}^{-2}$.

In Figure 2(right), we examine the relative errors for all the random pairs at a wide range of ℓ_2 norms, for $t = 100$ and $t = 1000$. A slight change in the error profile occurs within $\|x_i - y_i\|/\sigma \in [10^0, 10^1]$, coinciding with the separation of cases $\|x - y\| \leq \sigma$ and $\|x - y\| > \sigma$ i.e. whether $\frac{\|x-y\|}{\sigma} = \Theta(1)$ in the analyses.

In either case, the relative error is bounded by a small constant value, even when $\|x_i - y_i\|$ is several magnitudes smaller than 1, demonstrating that the extremely high concentration of the RFF for very small $\|x_i - y_i\|$ results in relative error approximation for the Gaussian kernel distance.

References

Amirali Abdullah, Ravi Kumar, Andrew McGregor, Sergei Vassilvitskii, and Suresh Venkatasubramanian. Sketching, embedding, and dimensionality reduction for information spaces. In *AIStats*, 2016.

Pankaj K. Agarwal and R. Sharathkumar. Approximation algorithms for bipartite matching with metric and geometric costs. In *STOC*, 2014.

Nir Ailon, Ragesh Jaiswal, and Claire Monteleoni. Streaming k-means approximation. In *NIPS*, 2009.

Noga Alon. Problems and results in extremal combinatorics-i. *Discrete Math.*, 273(1-3):31–53, 2003.

David Arthur and Sergei Vassilvitskii. k-means++: The advantage of careful seeding. In *SODA*, 2007.

Haim Avron, Huy L. Nguyen, and David P. Woodruff. Subspace embeddings for the polynomial kernel. In *NIPS*, 2014.

P. Massart B. Laurent. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. ISSN 00905364. URL http://www.jstor.org/stable/2674095.

Valerij Vladimirovič Buldygin, IU. V. Kozachenko, and V. Zaiats. *Metric characterization of random variables and random processes*. Translations of mathematical monographs. Providence, R.I. American Mathematical Society, 2000. ISBN 0-8218-0533-9. URL http://opac.inria.fr/record=b1132854. Traduit du russe : Metricheskie kharakteristiki sluchaĭnykh velichin i protsessov (1998).

Kenneth L. Clarkson and David P. Woodruff. Sketching for M-estimators: A unified approach to robust regression. In *SODA*, 2015.

Iderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means, spectral clustering and normalized cuts. In *KDD*, 2004.

Stanley Durrleman, Xavier Pennec, Alain Trouvé, and Nicholas Ayache. Measuring brain variability via sulcal lines registration: A diffeomorphic approach. In *10th International Conference on Medical Image Computing and Computer Assisted Intervention*, 2007.

Mark Girolami. Mercer kernel based clustering in feature space. *IEEE Transactions on Neural Networks*, 13:780–784, 2002.

Joan Glaunès and Sarang Joshi. Template estimation form unlabeled point set data and surfaces for computational anatomy. In *Math. Found. Comp. Anatomy*, 2006.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 1985.

Arthur Gretton, Marsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alex J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Raffay Hamid, Ying Xiao, Alex Gittens, and Dennis DeCoste. Compact random feature maps. In *ICML*, 2014.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Sarang Joshi, Raj Varma Kommaraju, Jeff M. Phillips, and Suresh Venkatasubramanian. Comparing distributions and shapes using the kurrent distance. In *Proceedings 27th Annual Symposium on Computational Geometry*, 2011. arXiv:1001.0591.

Purushottam Kar and Harish Karnick. Random feature maps for dot product kernels. *AISTATS*, 2012.

Kasper Green Larsen and Jelani Nelson. The johnson-lindenstrauss lemma is optimal for linear dimensionality reduction. In *ICALP*, 2016.

Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. In *FOCS*, 2017.

Fuxin Li, Catalin Ionescu, and Cristian Sminchisescu. Random fourier approximations for skewed multiplicative histogram kernels. In *Pattern Recognition*, pages 262–271. Springer, 2010.

David Lopez-Paz, Suvrit Sra, Alex Smola, Zoubin Ghahramani, and Bernhard Schölkopf. Randomized nonlinear component analysis. *ICML*, 2014.

Subhransu Maji and Alexander C Berg. Max-margin additive classifiers for detection. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 40–47. IEEE, 2009.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2007.

R. Sharathkumar and Pankaj K. Agarwal. A near-linear time ε-approximation algorithm for geometric bipartite matching. In *STOC*, 2012.

Alex J. Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *ICALT*, 2007.

Bharath K. Sriperumbudur and Zoltan Szabo. Optimal rates for random fourier features. In *NIPS*, 2015.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Dougal J. Sutherland and Jeff Schneider. On the error of random fourier features. In *UAI*, 2015.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005.

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

Appendix A. Omitted proof

Lemma 5[Restated] *For* $\omega \sim \mathcal{N}(0,1)$ *and* $0 \leq \|\Delta\| \leq 1$, *let* $M(s)$ *be the moment generating function of* $1 - \cos(\omega\|\Delta\|) - \left(1 - e^{-\frac{1}{2}\|\Delta\|^2}\right) = e^{-\frac{1}{2}\|\Delta\|^2} - \cos(\omega\|\Delta\|)$. *Then for all* $s \geq 0$,

$$\ln M(s) \leq \frac{1}{4}s^2\|\Delta\|^4.$$

Proof [Lemma 5] Recall that the moment generating function $M(s)$ of a random variable X is given by $\mathbf{E}[e^{sX}]$.

First we note two Taylor approximations which hold for all $x \in \mathbb{R}$:

$$\cos x \geq 1 - \frac{1}{2}x^2 \qquad \text{and} \qquad e^{-|x|} \leq 1 - |x| + \frac{1}{2}|x|^2.$$

Now

$$\begin{aligned} M(s) = & \ \mathbf{E}\left[\exp\left(s(e^{-\frac{1}{2}\|\Delta\|^2} - \cos(\omega\|\Delta\|))\right)\right] \\ \leq & \ \mathbf{E}\left[\exp\left(s(1-\frac{1}{2}\|\Delta\|^2 + \frac{1}{8}\|\Delta\|^4) - s(1-\frac{1}{2}\omega^2\|\Delta\|^2)\right)\right] \\ = & \ \mathbf{E}\left[\exp\left(-\frac{s}{2}\|\Delta\|^2 + \frac{s}{8}\|\Delta\|^4 + \frac{s}{2}\omega^2\|\Delta\|^2\right)\right] \\ = & \exp(\frac{s}{8}\cdot\|\Delta\|^4 - \frac{s}{2}\|\Delta\|^2)\cdot \ \mathbf{E}\left[e^{-s\frac{1}{2}\omega^2\|\Delta\|^2}\right]. \end{aligned}$$

But

$$\begin{aligned} \mathbf{E}\left[e^{-s\frac{1}{2}\omega^2\|\Delta\|^2}\right] &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}\cdot e^{-s\frac{1}{2}u^2\|\Delta\|^2}\mathrm{d}u \\ &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2(1+s\|\Delta\|^2)}\mathrm{d}u \\ &= \frac{1}{\sqrt{1+s\|\Delta\|^2}}\int_{-\infty}^{\infty}\sqrt{1+s\|\Delta\|^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2(1+s\|\Delta\|^2)}\mathrm{d}u \\ &= \frac{1}{\sqrt{1+s\|\Delta\|^2}}. \end{aligned}$$

Noting that $\ln(1+x) \geq x - \frac{x^2}{2}$ for $x \geq 0$, then whenever $s \geq 0$:

$$\begin{aligned} \ln M(s) &\leq \ln\left(\frac{\exp(\frac{s}{8}\|\Delta\|^4 - \frac{s}{2}\|\Delta\|^2)}{\sqrt{1+s\|\Delta\|^2}}\right) \\ &= \frac{s}{8}\|\Delta\|^4 - \frac{s}{2}\|\Delta\|^2 - \frac{1}{2}\ln(1+s\|\Delta\|^2) \\ &\leq \frac{s}{8}\|\Delta\|^4 - \frac{s}{2}\|\Delta\|^2 - \frac{1}{2}(s\|\Delta\|^2 - \frac{1}{2}s^2\|\Delta\|^4) \\ &= \frac{s^2}{4}\|\Delta\|^4 - \frac{s}{8}\|\Delta\|^4 - s\|\Delta\|^2 \\ &\leq \frac{s^2}{4}\|\Delta\|^4. \end{aligned}$$

■