On the Error of Random Fourier Features

Danica J. Sutherland
Carnegie Mellon University
Pittsburgh, PA
dsutherl@cs.cmu.edu

Jeff Schneider
Carnegie Mellon University
Pittsburgh, PA
schneide@cs.cmu.edu

Abstract

Kernel methods give powerful, flexible, and theoretically grounded approaches to solving many problems in machine learning. The standard approach, however, requires pairwise evaluations of a kernel function, which can lead to scalability issues for very large datasets. Rahimi and Recht (2007) suggested a popular approach to handling this problem, known as random Fourier features. The quality of this approximation, however, is not well understood. We improve the uniform error bound of that paper, as well as giving novel understandings of the embedding's variance, approximation error, and use in some machine learning methods. We also point out that surprisingly, of the two main variants of those features, the more widely used is strictly higher-variance for the Gaussian kernel and has worse bounds.

1 INTRODUCTION

Kernel methods provide an elegant, theoretically well-founded, and powerful approach to solving many learning problems. Since traditional algorithms require the computation of a full $N \times N$ pairwise kernel matrix to solve learning problems on N input instances, however, scaling these methods to large-scale datasets containing more than thousands of data points has proved challenging. Rahimi and Recht (2007) spurred interest in one very attractive approach: approximating a continuous shift-invariant kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$k(x, y) \approx z(x)^\mathsf{T} z(y) =: s(x, y),$$

where $z : \mathcal{X} \to \mathbb{R}^D$. Then primal methods in $\mathbb{R}^D$ can be used, allowing most learning problems to be solved in $O(N)$ time (e.g. Joachims 2006). Recent work has also exploited these embeddings in some of the most-scalable kernel methods to date (Dai et al. 2014).

Rahimi and Recht (2007) give two such embeddings, based on the Fourier transform $P(\omega)$ of the kernel k: one of the form

$$\tilde{z}(x) := \sqrt{\frac{2}{D}} \begin{bmatrix} \sin(\omega_1^\mathsf{T} x) \\ \cos(\omega_1^\mathsf{T} x) \\ \vdots \\ \sin(\omega_{D/2}^\mathsf{T} x) \\ \cos(\omega_{D/2}^\mathsf{T} x) \end{bmatrix}, \; \omega_i \overset{iid}{\sim} P(\omega) \tag{1}$$

and another of the form

$$\breve{z}(x) := \sqrt{\frac{2}{D}} \begin{bmatrix} \cos(\omega_1^\mathsf{T} x + b_1) \\ \vdots \\ \cos(\omega_D^\mathsf{T} x + b_D) \end{bmatrix}, \; \begin{matrix} \omega_i \overset{iid}{\sim} P(\omega) \\ b_i \overset{iid}{\sim} \mathrm{Unif}_{[0,2\pi]} \end{matrix}. \tag{2}$$

Bochner's theorem (1959) guarantees that for any continuous positive-definite function $k(x-y)$, its Fourier transform will be a nonnegative measure; if $k(0) = 1$, it will be properly normalized. Letting $\tilde{s}$ be the reconstruction based on $\tilde{z}$ and $\breve{s}$ that for $\breve{z}$, we have that:

$$\tilde{s}(x, y) = \frac{1}{D/2} \sum_{i=1}^{D/2} \cos(\omega_i^\mathsf{T}(x - y))$$

$$\breve{s}(x, y) = \frac{1}{D} \sum_{i=1}^{D} \cos(\omega_i^\mathsf{T}(x - y)) + \cos(\omega_i^\mathsf{T}(x + y) + 2b_i).$$

Letting $\Delta := x - y$, we have:

$$\mathbb{E} \cos(\omega^\mathsf{T} \Delta) = \Re \int e^{\omega^\mathsf{T} \Delta \sqrt{-1}} \mathrm{d}P(\omega) = \Re k(\Delta) \tag{3}$$

$$\mathbb{E}_\omega \mathbb{E}_b \cos(\omega^\mathsf{T}(x + y) + 2b) = 0. \tag{4}$$

Thus each $s(x, y)$ is a mean of bounded terms with expectation $k(x, y)$. For a given embedding dimension D, it is not immediately obvious which approximation is preferable: $\breve{z}$ gives twice as many samples for ω, but adds additional (non-shift-invariant) noise. The academic literature seems split on the issue: of the first 100 papers citing Rahimi and Recht (2007) in a Google Scholar search, 15 used either $\tilde{z}$ or the equivalent complex formulation, 14 used $\breve{z}$, 28 did not specify, and the remainder didn't use the embedding. (None discussed that there was a choice.) Not included in the count are are Rahimi and Recht's later work (2008a; 2008b), which used $\breve{z}$; indeed, post-publication revisions of the original paper only discuss $\breve{z}$. Practically, we are aware of three implementations in machine learning libraries, each of which use $\breve{z}$ at the time of writing: scikit-learn (Pedregosa et al. 2011), Shogun (Sonnenburg et al. 2010), and JSAT (Raff 2011-15).

We show that $\tilde{z}$ is superior for the popular Gaussian kernel, as well as how to decide which to use for other kernels.

The primary previous analyses of these embeddings, outside the one in the original paper, have been by Rahimi and Recht (2008a), who bound the increase in error of empirical risk estimates when learning models in the induced RKHS, and by Yang et al. (2012), who compare the ability of the Nyström and Fourier embeddings to exploit eigengaps in the learning problem. We instead study the approximation directly, providing a complementary view of the quality of these embeddings.

Section 2.1 studies the variance of each embedding, showing that which is preferable depends on the kernel as well as the particular value of Δ, but for the popular Gaussian kernel $\tilde{s}$ is uniformly lower-variance. Section 2.2 studies uniform convergence bounds, tightening constants in the original $\tilde{z}$ bound and proving a comparable one (with worse constants) for $\breve{z}$, bounding the expectation of the maximal error, and providing exponential concentration about the mean. Section 2.3 studies the L_2 convergence of each approximation; $\tilde{z}$ is again superior for the Gaussian kernel. Section 3 discusses the effect of this approximation error when used in various machine learning methods. Section 4 evaluates the two embeddings and the bounds empirically.

2 APPROXIMATION ERROR

We will give various analyses of the error due to each approximation.

2.1 VARIANCE

(3) and (4) establish that $\mathbb{E}s(\Delta) = k(\Delta)$. What about the variance? We have that

$$\begin{aligned}
&\operatorname{Cov}\left(\tilde{s}(\Delta), \tilde{s}(\Delta')\right) \\
&\quad= \operatorname{Cov}\left(\frac{2}{D} \sum_{i=1}^{D/2} \cos(\omega_i^\mathsf{T} \Delta), \frac{2}{D} \sum_{i=1}^{D/2} \cos(\omega_i^\mathsf{T} \Delta') \right) \\
&\quad= \frac{2}{D} \operatorname{Cov}\left(\cos(\omega^\mathsf{T}\Delta), \cos(\omega^\mathsf{T}\Delta')\right) \\
&\quad= \frac{2}{D} \left[\tfrac{1}{2} k(\Delta - \Delta') + \tfrac{1}{2} k(\Delta + \Delta') - k(\Delta) k(\Delta') \right]
\end{aligned}$$

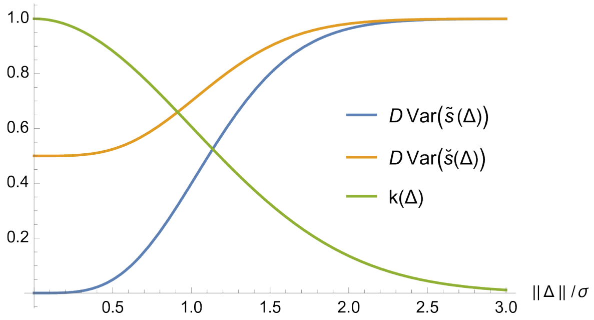

Figure 1: The variance per dimension of $\tilde{s}$ (blue) and $\breve{s}$ (orange) for the Gaussian RBF kernel (green).

using $\cos(\alpha)\cos(\beta) = \frac{1}{2}\cos(\alpha+\beta) + \frac{1}{2}\cos(\alpha-\beta)$ and also $\mathbb{E}\cos(\omega^\mathsf{T}\Delta) = k(\Delta)$. Thus

$$\operatorname{Var} \tilde{s}(\Delta) = \frac{1}{D}\left[1 + k(2\Delta) - 2k(\Delta)^2\right]. \tag{5}$$

Similarly, denoting $x + y$ by t,

$$\begin{aligned}
&\operatorname{Cov}\left(\breve{s}(x,y), \breve{s}(x',y')\right) \\
&\quad= \frac{1}{D} \operatorname{Cov}\big(\cos(\omega^\mathsf{T}\Delta) + \cos(\omega^\mathsf{T} t + 2b), \\
&\qquad\qquad \cos(\omega^\mathsf{T}\Delta') + \cos(\omega^\mathsf{T} t' + 2b)\big) \\
&\quad= \frac{1}{D} \big[\tfrac{1}{2} k(\Delta - \Delta') + \tfrac{1}{2} k(\Delta + \Delta') - k(\Delta) k(\Delta') \\
&\qquad + \tfrac{1}{2} k(t - t') \big]
\end{aligned}$$

which gives

$$\operatorname{Var} \breve{s}(x,y) = \frac{1}{D}\left[1 + \tfrac{1}{2}k(2\Delta) - k(\Delta)^2\right]. \tag{6}$$

Thus $\tilde{s}$ has lower variance than $\breve{s}$ if

$$\operatorname{Var}\cos(\omega^\mathsf{T}\Delta) = \frac{1}{2} + \frac{1}{2}k(2\Delta) - k(\Delta)^2 \le \frac{1}{2}. \tag{7}$$

The Gaussian kernel $k(\Delta) = \exp\left(-\frac{\lVert\Delta\rVert^2}{2\sigma^2}\right)$ has

$$\operatorname{Var}\cos(\omega^\mathsf{T}\Delta) = \frac{1}{2}\left(1 - \exp\left(-\frac{\lVert\Delta\rVert^2}{\sigma^2}\right)\right)^2 \le \frac{1}{2},$$

so that $\tilde{z}$ is always lower-variance than $\breve{z}$, and the difference in variance is greatest when $k(\Delta)$ is largest. This is illustrated in Figure 1.

2.2 UNIFORM ERROR BOUND

Let $f(x, y) := s(x, y) - k(x, y)$ denote the error of the approximation. We will investigate $\lVert f \rVert_\infty$, i.e. the maximal approximation error across the domain of k. We first consider the bound given by Rahimi and Recht (2007), and then provide a new bound on $\mathbb{E}\lVert f\rVert_\infty$ and its concentration around that mean.

2.2.1 Original High-Probability Bound

Claim 1 of Rahimi and Recht (2007) is that if $\mathcal{X} \subset \mathbb{R}^d$ is compact with diameter ℓ,[1]

$$\Pr\left(\|f\|_\infty \geq \varepsilon\right) \leq 256 \left(\frac{\sigma_p \ell}{\varepsilon}\right)^2 \exp\left(-\frac{D\varepsilon^2}{8(d+2)}\right),$$

where $\sigma_p^2 = \mathbb{E}\left[\omega^\mathsf{T}\omega\right] = \operatorname{tr} \nabla^2 k(0)$ depends on the kernel.

It is not necessarily clear in that paper that this bound applies only to the $\tilde{z}$ embedding; we can also tighten some constants. We first state the tightened bound for $\tilde{z}$.

Proposition 1. *Let k be a continuous shift-invariant positive-definite function $k(x, y) = k(\Delta)$ defined on $\mathcal{X} \subset \mathbb{R}^d$, with $k(0) = 1$ and such that $\nabla^2 k(0)$ exists. Suppose $\mathcal{X}$ is compact, with diameter ℓ. Denote k's Fourier transform as $P(\omega)$, which will be a probability distribution; let $\sigma_p^2 = \mathbb{E}_p \|\omega\|^2$. Let $\tilde{z}$ be as in (1), and define $\tilde{f}(x, y) := \tilde{z}(x)^\mathsf{T}\tilde{z}(y) - k(x, y)$. For any $\varepsilon > 0$, let*

$$\alpha_\varepsilon := \min\left(1, \sup_{x,y\in\mathcal{X}} \frac{1}{2} + \frac{1}{2}k(2x, 2y) - k(x, y)^2 + \tfrac{1}{3}\varepsilon\right),$$

$$\beta_d := \left(\left(\tfrac{d}{2}\right)^{\frac{-d}{d+2}} + \left(\tfrac{d}{2}\right)^{\frac{2}{d+2}}\right) 2^{\frac{6d+2}{d+2}}.$$

Then, assuming only for the second statement that $\varepsilon \leq \sigma_p \ell$,

$$\begin{aligned}\Pr\left(\|\tilde{f}\|_\infty \geq \varepsilon\right) &\leq \beta_d \left(\frac{\sigma_p \ell}{\varepsilon}\right)^{\frac{2}{1+\frac{2}{d}}} \exp\left(-\frac{D\varepsilon^2}{8(d+2)\alpha_\varepsilon}\right) \\ &\leq 66 \left(\frac{\sigma_p \ell}{\varepsilon}\right)^2 \exp\left(-\frac{D\varepsilon^2}{8(d+2)}\right).\end{aligned}$$

Thus, we can achieve an embedding with pointwise error no more than ε with probability at least $1 - \delta$ as long as

$$D \geq \frac{8(d+2)\alpha_\varepsilon}{\varepsilon^2}\left[\frac{2}{1+\frac{2}{d}} \log \frac{\sigma_p \ell}{\varepsilon} + \log \frac{\beta_d}{\delta}\right].$$

The proof strategy is very similar to that of Rahimi and Recht (2007): place an ε-net with radius r over $\mathcal{X}_\Delta := \{x - y : x, y \in \mathcal{X}\}$, bound the error $\tilde{f}$ by $\varepsilon/2$ at the centers of the net by Hoeffding's inequality (1963), and bound the Lipschitz constant of $\tilde{f}$, which is at most that of $\tilde{s}$, by $\varepsilon/(2r)$ with Markov's inequality. The introduction of α_ε is by replacing Hoeffding's inequality with that of Bernstein (1924) when it is tighter, using the variance from (5). The constant β_d is obtained by exactly optimizing the value of r, rather than the algebraically simpler value originally used; $\beta_{64} = 66$ is its maximum, and $\lim_{d\to\infty} \beta_d = 64$, though it is lower for small d, as shown in Figure 2. The additional hypothesis, that $\nabla^2 k(0)$ exists, is equivalent to the existence of the first two moments of $P(\omega)$; a finite first moment is used in the proof, and of course without a finite second moment the bound is vacuous. The full proof is given in Appendix A.1.

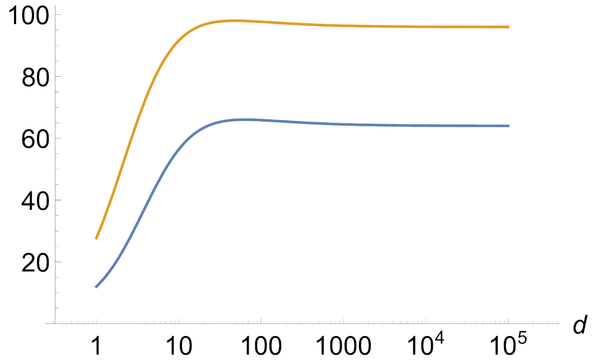

Figure 2: The coefficient β_d of Proposition 1 (blue, for $\tilde{z}$) and β'_d of Proposition 2 (orange, for $\breve{z}$). Rahimi and Recht (2007) used a constant of 256 for $\tilde{z}$.

For the Gaussian kernel, $\alpha_\varepsilon \leq \frac{1}{2} + \frac{1}{3}\varepsilon$ and $\sigma_p^2 = d/\sigma^2$; the Bernstein bound is tighter when $\varepsilon < \frac{3}{2}$.

For $\breve{z}$, since the embedding $\breve{s}$ is not shift-invariant, we must instead place the ε-net on $\mathcal{X}^2$. The additional noise in $\breve{s}$ also increases the expected Lipschitz constant and gives looser bounds on each term in the sum, though there are twice as many such terms. The corresponding bound is as follows:

Proposition 2. *Let k, $\mathcal{X}$, ℓ, $P(\omega)$, and σ_p be as in Proposition 1. Define $\breve{z}$ by (2), and $\breve{f}(x, y) := \breve{z}(x)^\mathsf{T}\breve{z}(y) - k(x, y)$. For any $\varepsilon > 0$, define*

$$\alpha'_\varepsilon := \min\left(1, \sup_{x,y\in\mathcal{X}} \tfrac{1}{4} + \tfrac{1}{8}k(2x, 2y) - \tfrac{1}{4}k(x, y)^2 + \tfrac{1}{6}\varepsilon\right),$$

$$\beta'_d := \left(d^{\frac{-d}{d+1}} + d^{\frac{1}{d+1}}\right) 2^{\frac{5d+1}{d+1}} 3^{\frac{d}{d+1}}.$$

Then, assuming only for the second statement that $\varepsilon \leq \sigma_p \ell$,

$$\begin{aligned}\Pr\left(\|\breve{f}\|_\infty \geq \varepsilon\right) &\leq \beta'_d \left(\frac{\sigma_p \ell}{\varepsilon}\right)^{\frac{2}{1+\frac{1}{d}}} \exp\left(-\frac{D\varepsilon^2}{32(d+1)\alpha'_\varepsilon}\right) \\ &\leq 98 \left(\frac{\sigma_p \ell}{\varepsilon}\right)^2 \exp\left(-\frac{D\varepsilon^2}{32(d+1)}\right).\end{aligned}$$

Thus, we can achieve an embedding with pointwise error no more than ε with probability at least $1 - \delta$ as long as

$$D \geq \frac{32(d+1)\alpha'_\varepsilon}{\varepsilon^2}\left[\frac{2}{1+\frac{1}{d}} \log \frac{\sigma_p \ell}{\varepsilon} + \log \frac{\beta'_d}{\delta}\right].$$

$\beta'_{48} = 98$, and $\lim_{d\to\infty} \beta'_d = 96$, also shown in Figure 2. The full proof is given in Appendix A.2.

For the Gaussian kernel, $\alpha'_\varepsilon \leq \frac{1}{4} + \frac{1}{6}\varepsilon$, so that the Berstein bound is essentially always superior.

2.2.2 Expected Max Error

Noting that $\mathbb{E}\|f\|_\infty = \int_0^\infty \Pr\left(\|f\|_\infty \geq \varepsilon\right) \mathrm{d}\varepsilon$, one could consider bounding $\mathbb{E}\|f\|_\infty$ via Propositions 1 and 2. Unfortunately, that integral diverges on $(0, \gamma)$ for any $\gamma > 0$.

[1] Note that our D is half of the D in Rahimi and Recht (2007), since we want to compare embeddings of the same dimension.

If we instead integrate the minimum of that bound and 1, the result depends on a solution to a transcendental equation, so analytical manipulation is difficult.

We can, however, use a slight generalization of Dudley's entropy integral (1967) to obtain the following bound:

Proposition 3. *Let k, $\mathcal{X}$, ℓ, and $P(\omega)$ be as in Proposition 1. Define $\tilde z$ by (1), and $\tilde f(x,y) := \tilde z(x)^\mathsf{T} \tilde z(y) - k(x,y)$. Let $\mathcal{X}_\Delta := \{x - y \mid x, y \in \mathcal{X}\}$; suppose k is L-Lipschitz on $\mathcal{X}_\Delta$. Let $R := \mathbb{E}\max_{i=1,\dots,\frac{D}{2}} \|\omega_i\|$. Then*

$$\mathbb{E}\left[\|\tilde f\|_\infty\right] \le \frac{24\gamma\sqrt{d}\ell}{\sqrt{D}}(R+L)$$

where $\gamma \approx 0.964$.

The proof is given in Appendix A.3. In order to apply the method of Dudley (1967), we must work around $\|\omega_i\|$ (which appears in the covariance of the error process) being potentially unbounded. To do so, we bound a process with truncated $\|\omega_i\|$, and then relate that bound to $\tilde f$.

For the Gaussian kernel, $L = 1/(\sigma\sqrt{e})$ and[2]

$$\begin{aligned} R &\le \left(\sqrt{2}\,\frac{\Gamma\left((d+1)/2\right)}{\Gamma\left(d/2\right)} + \sqrt{2\log\left(D/2\right)}\right)/\sigma \\ &\le \left(\sqrt{d} + \sqrt{2\log\left(D/2\right)}\right)/\sigma. \end{aligned}$$

Thus $\mathbb{E}\|\tilde f\|_\infty$ is less than

$$\frac{24\gamma\sqrt{d}\,\ell}{\sqrt{D}\,\sigma}\left(e^{-1/2} + \sqrt{d} + \sqrt{2\log(D/2)}\right). \qquad (8)$$

We can also prove an analogous bound for the $\breve z$ features:

Proposition 4. *Let $k, \mathcal{X}, \ell$, and $P(\omega)$ be as in Proposition 1. Define $\breve z$ by (2), and $\breve f(x,y) := \breve z(x)^\mathsf{T}\breve z(y) - k(x,y)$. Suppose $k(\Delta)$ is L-Lipschitz. Let $R := \mathbb{E}\max_{i=1,\dots,D}\|\omega_i\|$. Then, for $\mathcal{X}$ and D not extremely small,*

$$\mathbb{E}\left[\|\breve f\|_\infty\right] \le \frac{48\gamma'_{\mathcal{X}}\ell\sqrt{d}}{\sqrt{D}}(R+L)$$

where $0.803 < \gamma'_{\mathcal{X}} < 1.542$. See Appendix A.4 for details on $\gamma'_{\mathcal{X}}$ and the "not extremely small" assumption.

The proof is given in Appendix A.4. It is similar to that for Proposition 3, but the lack of shift invariance increases some constants and otherwise slightly complicates matters. Note also that the R of Proposition 4 is somewhat larger than that of Proposition 3.

2.2.3 Concentration About Mean

Bousquet's inequality (2002) can be used to show exponential concentration of $\sup f$ about its mean.

We consider $\tilde f$ first. Let

$$\tilde f_\omega(\Delta) := \frac{2}{D}\left(\cos(\omega^\mathsf{T}\Delta) - k(\Delta)\right),$$

so $f(\Delta) = \sum_{i=1}^{D/2} \tilde f_{\omega_i}(\Delta)$. Define the "wimpy variance" of $\tilde f/2$ (which we use so that $|\tilde f/2| \le 1$) as

$$\begin{aligned} \sigma^2_{\tilde f/2} :&= \sup_{\Delta\in\mathcal{X}_\Delta} \sum_{i=1}^{D/2} \operatorname{Var}\left[\tfrac{1}{2}\tilde f_{\omega_i}(\Delta)\right] \\ &= \frac{1}{D}\sup_{\Delta\in\mathcal{X}_\Delta}\left[1 + k(2\Delta) - 2k(\Delta)^2\right] \\ &=: \frac{1}{D}\sigma_w^2, \end{aligned}$$

using (7). Clearly $1 \le \sigma_w^2 \le 2$; for the Gaussian kernel, it is 1.

Proposition 5. *Let k, $\mathcal{X}$, and $P(\omega)$ be as in Proposition 1, and $\tilde z$ be defined by (1). Let $\tilde f(\Delta) = \tilde z(x)^\mathsf{T}\tilde z(y) - k(\Delta)$ for $\Delta = x - y$, and $\sigma_w^2 := \sup_{\Delta\in\mathcal{X}_\Delta} 1 + k(2\Delta) - 2k(\Delta)^2$. Then*

$$\begin{aligned} &\Pr\left(\|\tilde f\|_\infty - \mathbb{E}\|\tilde f\|_\infty \ge \varepsilon\right) \\ &\qquad \le 2\exp\left(-\frac{D\varepsilon^2}{D\,\mathbb{E}\|\tilde f\|_\infty + \frac{1}{2}\sigma_w^2 + \frac{D\varepsilon}{6}}\right). \end{aligned}$$

Proof. We use the Bernstein-style form of Theorem 12.5 of Boucheron et al. (2013) on $\tilde f(\Delta)/2$ to obtain that $\Pr\left(\sup\tilde f - \mathbb{E}\sup\tilde f \ge \varepsilon\right)$ is at most

$$\exp\left(-\frac{\varepsilon^2}{\mathbb{E}\sup\tilde f + \frac{1}{2}\sigma^2_{\tilde f/2} + \frac{\varepsilon}{6}}\right).$$

The same holds for $-\tilde f$, and $\mathbb{E}\sup\tilde f \le \mathbb{E}\sup\|f\|_\infty$, $\mathbb{E}\sup(-\tilde f) \le \mathbb{E}\sup\|f\|_\infty$. The claim follows by a union bound. □

A bound on the lower tail, unfortunately, is not available in the same form.

For $\breve f$, note $|\breve f| \le 3$, so we use $\breve f/3$. Letting $\breve f_\omega := \frac{1}{D}(\cos(\omega^\mathsf{T}\Delta) - k(\Delta))$, we have $\sigma^2_{\breve f/3} = \frac{1}{18D}(\sigma_w^2 + 1)$. Thus the same argument gives us:

Proposition 6. *Let k and $\mathcal{X}$ be as in Proposition 1, with $P(\omega)$ defined as there. Let $\breve z$ be as in (2), $\breve f(x,y) = \breve z(x)^\mathsf{T}\breve z(y) - k(x,y)$, and define σ_w as above. Then*

$$\begin{aligned} &\Pr\left(\|\breve f\|_\infty - \mathbb{E}\|\breve f\|_\infty \ge \varepsilon\right) \\ &\quad \le 2\exp\left(-\frac{D\varepsilon^2}{\frac{4}{9}D\,\mathbb{E}\|\breve f\|_\infty + \frac{1}{81}(\sigma_w^2+1) + \frac{2}{27}D\varepsilon}\right). \end{aligned}$$

[2] By the Gaussian concentration inequality (Boucheron et al. 2013, Theorem 5.6), each $\|\omega\| - \mathbb{E}\|\omega\|$ is sub-Gaussian with variance factor σ^{-2}; the claim follows from their Section 2.5.

Note that Proposition 6 actually gives a somewhat tighter concentration than Proposition 5. This is most likely because, between the space of possible errors being larger and the higher base variance illustrated in Figure 1, the $\breve{f}$ error function has more "opportunities" to achieve its maximal error. The experimental results (Figure 5) show that, at least in one case, $\|\breve{f}\|_\infty$ does concentrate about its mean more tightly, but that mean is enough higher than that of $\|\tilde{f}\|_\infty$ that $\|\breve{f}\|_\infty$ stochastically dominates $\|\tilde{f}\|_\infty$.

2.3 L_2 ERROR BOUND

L_∞ bounds provide useful guarantees, but are very strict. It can also be useful to consider a less stringent error measure. Let μ be a σ-finite measure on $\mathcal{X} \times \mathcal{X}$; define

$$\|f\|_\mu^2 := \int_{\mathcal{X}^2} f(x,y)^2 \, \mathrm{d}\mu(x,y). \tag{9}$$

First, we have that

$$\begin{aligned}
\mathbb{E}\|\tilde{f}\|_\mu^2 &= \mathbb{E} \int_{\mathcal{X}^2} \tilde{f}(x,y)^2 \, \mathrm{d}\mu(x,y) \\
&= \int_{\mathcal{X}^2} \mathbb{E}\, \tilde{f}(x,y)^2 \, \mathrm{d}\mu(x,y) \qquad (10) \\
&= \int_{\mathcal{X}^2} \frac{1}{D} \left[1 + k(2x,2y) - 2k(x,y)^2 \right] \mathrm{d}\mu(x,y) \\
&= \frac{1}{D} \left[\mu(\mathcal{X}^2) + \int_{\mathcal{X}^2} k(2x,2y) \, \mathrm{d}\mu(x,y) - 2\, \|k\|_\mu^2 \right] \\
\mathbb{E}\|\breve{f}\|_\mu^2 &= \frac{1}{D} \left[\mu(\mathcal{X}^2) + \frac{1}{2} \int_{\mathcal{X}^2} k(2x,2y) \, \mathrm{d}\mu(x,y) - \|k\|_\mu^2 \right]
\end{aligned}$$

where (10) is justified by Tonelli's theorem.

If $\mu = P_X \times P_Y$ is a joint distribution of independent variables, then $\int_{\mathcal{X}^2} k(2x,2y) \, \mathrm{d}\mu(x,y) = \mathrm{MMK}(P_{2X}, P_{2Y})$, where MMK is the mean map kernel (see Section 3.3). Likewise, $\|k\|_\mu^2 = \mathrm{MMK}(P_X, P_Y)$ using the kernel k^2.[3]

Viewing $\|\tilde{f}\|_\mu$ as a function of $\omega_1, \dots, \omega_{D/2}$, changing ω_i to a different $\hat{\omega}_i$ changes the value of $\|\tilde{f}\|_\mu$ by at most $4 \frac{4D+1}{D^2} \mu(\mathcal{X}^2)$; this can be seen by simple algebra and is shown in Appendix B.1. Thus McDiarmid (1989) gives us an exponential concentration bound:

Proposition 7. *Let k be a continuous shift-invariant positive-definite function $k(x,y) = k(\Delta)$ defined on $\mathcal{X} \subseteq \mathbb{R}^d$, with $k(0) = 1$. Let μ be a σ-finite measure on $\mathcal{X}^2$, and define $\|\cdot\|_\mu^2$ as in (9). Define $\tilde{z}$ as in (1) and let $\tilde{f}(x,y) = \tilde{z}(x)^\mathsf{T} \tilde{z}(y) - k(x,y)$. Let $\mathcal{M} := \mu(\mathcal{X}^2)$. Then*

$$\begin{aligned}
\Pr\left(\left| \|\tilde{f}\|_\mu^2 - \mathbb{E}\|\tilde{f}\|_\mu^2 \right| \geq \varepsilon \right) &\leq 2 \exp\left(\frac{-D^3 \varepsilon^2}{8(4D+1)^2 \, \mathcal{M}^2} \right) \\
&\leq 2 \exp\left(\frac{-D\varepsilon^2}{200\, \mathcal{M}^2} \right).
\end{aligned}$$

The second version of the bound is simpler, but somewhat looser for $D \gg 1$; asymptotically, the coefficient of the denominator becomes 128.

Similarly, the variation of $\|\breve{f}\|_\mu$ is bounded by at most $32 \frac{D+1}{D^2} \mu(\mathcal{X}^2)$ (shown in Appendix B.2). Thus:

Proposition 8. *Let k, μ, $\|\cdot\|_\mu$, and $\mathcal{M}$ be as in Proposition 7. Define $\breve{z}$ as in (2) and let $\breve{f}(x,y) = \breve{z}(x)^\mathsf{T} \breve{z}(y) - k(x,y)$. Then*

$$\begin{aligned}
\Pr\left(\left| \|\breve{f}\|_\mu^2 - \mathbb{E}\|\breve{f}\|_\mu^2 \right| \geq \varepsilon \right) &\leq 2 \exp\left(\frac{-D^3 \varepsilon^2}{512(D+1)^2 \, \mathcal{M}^2} \right) \\
&\leq 2 \exp\left(\frac{-D\varepsilon^2}{2048\, \mathcal{M}^2} \right).
\end{aligned}$$

The cost of a simpler dependence on D is higher here; the asymptotic coefficient of the denominator is 512.

3 DOWNSTREAM ERROR

Rahimi and Recht (2008a; 2008b) give a bound on the L_2 distance between any given function in the reproducing kernel Hilbert space (RKHS) induced by k and the closest function in the RKHS of s: results invaluable for the study of learning rates. In some situations, however, it is useful to consider not the learning-theoretic convergence of hypotheses to the assumed "true" function, but rather directly consider the difference in predictions due to using the z embedding instead of the exact kernel k.

3.1 KERNEL RIDGE REGRESSION

We first consider kernel ridge regression (KRR; Saunders et al. 1998). Suppose we are given n training pairs $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ as well as a regularization parameter $\lambda = n\lambda_0 > 0$. We construct the training Gram matrix K by $K_{ij} = k(x_i, x_j)$. KRR gives predictions $h(x) = \alpha^\mathsf{T} k_x$, where $\alpha = (K + \lambda I)^{-1} y$ and k_x is the vector with ith component $k(x_i, x)$.[4] When using Fourier features, one would not use α, but instead a primal weight vector w; still, it will be useful for us to analyze the situation in the dual.

Proposition 1 of Cortes et al. (2010) bounds the change in KRR predictions from approximating the kernel matrix K by $\hat{K}$, in terms of $\|\hat{K} - K\|_2$. They assume, however, that the kernel evaluations at test time k_x are unapproximated, which is certainly not the case when using Fourier features. We therefore extend their result to Proposition 9 before using it to analyze the performance of Fourier features.

[3] k^2 is also a PSD kernel, by the Schur product theorem.

[4] If a bias term is desired, we can use $k'(x,x') = k(x,x') + 1$ by appending a constant feature 1 to the embedding z. Because this change is accounted for exactly, it affects the error analysis here only in that we must use $\sup|k(x,y)| \leq 2$, in which case the first factor of (11) becomes $(\lambda_0 + 2)/\lambda_0^2$.

Proposition 9. *Given a training set $\{(x_i, y_i)\}_{i=1}^n$, with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, let $h(x)$ denote the result of kernel ridge regression using the* PSD *training kernel matrix K and test kernel values k_x. Let $\hat{h}(x)$ be the same using a* PSD *approximation to the training kernel matrix $\hat{K}$ and test kernel values $\hat{k}_x$. Further assume that the training labels are centered, $\sum_{i=1}^n y_i = 0$, and let $\sigma_y^2 := \frac{1}{n}\sum_{i=1}^n y_i^2$. Also suppose $\|k_x\|_\infty \le \kappa$. Then:*

$$|h'(x) - h(x)| \le \frac{\sigma_y}{\sqrt{n}\lambda_0}\|\hat{k}_x - k_x\| + \frac{\kappa\sigma_y}{n\lambda_0^2}\|\hat{K} - K\|_2.$$

Proof. Let $\alpha = (K + \lambda I)^{-1}y$, $\hat{\alpha} = (\hat{K} + \lambda I)^{-1}y$. Thus, using $\hat{M}^{-1} - M^{-1} = -\hat{M}^{-1}(\hat{M} - M)M^{-1}$, we have

$$\begin{aligned}\hat{\alpha} - \alpha &= -(\hat{K} + \lambda I)^{-1}(\hat{K} - K)(K + \lambda I)^{-1}y \\ \|\hat{\alpha} - \alpha\| &\le \|(\hat{K} + \lambda I)^{-1}\|_2\|\hat{K} - K\|_2\|(K + \lambda I)^{-1}\|_2\|y\| \\ &\le \frac{1}{\lambda^2}\|\hat{K} - K\|_2 \, \|y\|\end{aligned}$$

since the smallest eigenvalues of $K + \lambda I$ and $\hat{K} + \lambda I$ are at least λ. Since $\|k_x\| \le \sqrt{n}\kappa$ and $\|\hat{\alpha}\| \le \|y\|/\lambda$:

$$\begin{aligned}|\hat{h}(x) - h(x)| &= |\hat{\alpha}^\mathsf{T}\hat{k}_x - \alpha^\mathsf{T}k_x| \\ &= |\hat{\alpha}^\mathsf{T}(\hat{k}_x - k_x) + (\hat{\alpha} - \alpha)^\mathsf{T}k_x| \\ &\le \|\hat{\alpha}\|\|\hat{k}_x - k_x\| + \|\hat{\alpha} - \alpha\|\|k_x\| \\ &\le \frac{\|y\|}{\lambda}\|\hat{k}_x - k_x\| + \frac{\sqrt{n}\kappa\|y\|}{\lambda^2}\|\hat{K} - K\|_2.\end{aligned}$$

The claim follows from $\lambda = n\lambda_0$, $\|y\| = \sqrt{n}\sigma_y$. □

Suppose that, per the uniform error bounds of Section 2.2, $\sup|k(x,y) - s(x,y)| \le \varepsilon$. Then $\|\hat{k}_x - k_x\| \le \sqrt{n}\varepsilon$ and $\|\hat{K} - K\|_2 \le \|\hat{K} - K\|_F \le n\varepsilon$, and Proposition 9 gives

$$\begin{aligned}\left|\hat{h}(x) - h(x)\right| &\le \frac{\sigma_y}{\sqrt{n}\lambda_0}\sqrt{n}\varepsilon + \frac{\sigma_y}{n\lambda_0^2}n\varepsilon \\ &\le \frac{\lambda_0 + 1}{\lambda_0^2}\sigma_y\varepsilon. \qquad (11)\end{aligned}$$

Thus

$$\Pr\left(|h'(x) - h(x)| \ge \varepsilon\right) \le \Pr\left(\|f\|_\infty \ge \frac{\lambda_0^2\varepsilon}{(\lambda_0 + 1)\sigma_y}\right).$$

which we can bound with Proposition 1 or 2. We can therefore guarantee $|h(x) - h'(x)| \le \varepsilon$ with probability at least δ if

$$D = \Omega\left(d\left(\frac{(\lambda_0 + 1)\sigma_y}{\lambda_0^2\,\varepsilon}\right)^2 \left[\log\delta + \log\frac{\lambda_0^2\varepsilon}{(\lambda_0 + 1)\sigma_y} - \log\sigma_p\ell\right]\right).$$

Note that this rate does not depend on n. If we want $h'(x) \to h(x)$ at least as fast as $h(x)$'s convergence rate of $O(1/\sqrt{n})$ (Bousquet and Elisseeff 2001), ignoring the logarithmic terms, we thus need D to be linear in n, matching the conclusion of Rahimi and Recht (2008a).

3.2 SUPPORT VECTOR MACHINES

Consider a Support Vector Machine (SVM) classifier with no offset, such that $h(x) = w^\mathsf{T}\Phi(x)$ for a kernel embedding $\Phi(x) : \mathcal{X} \to \mathcal{H}$ and w is found by

$$\operatorname*{argmin}_{w \in \mathcal{H}} \frac{1}{2}\|w\|^2 + \frac{C_0}{n}\sum_{i=1}^n \max\left(0, 1 - y_i\langle w, \Phi(x_i)\rangle\right)$$

where $\{(x_i, y_i)\}_{i=1}^n$ is our training set with $y_i \in \{-1, 1\}$, and the decision function is $h(x) = \langle w, \Phi(x)\rangle$.[5] For a given x, Cortes et al. (2010) consider an embedding in $\mathcal{H} = \mathbb{R}^{n+1}$ which is equivalent on the given set of points. They bound $\left|\hat{h}(x) - h(x)\right|$ in terms of $\|\hat{K} - K\|_2$ in their Proposition 2, but again assume that the test-time kernel values k_x are exact. We will again extend their result in Proposition 10:

Proposition 10. *Given a training set $\{(x_i, y_i)\}_{i=1}^n$, with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$, let $h(x)$ denote the decision function of an* SVM *classifier using the* PSD *training matrix K and test kernel values k_x. Let $\hat{h}(x)$ be the same using a* PSD *approximation to the training kernel matrix $\hat{K}$ and test kernel values $\hat{k}_x$. Suppose $\sup k(x,x) \le \kappa$. Then:*

$$\begin{aligned}&|\hat{h}(x) - h(x)| \\ &\quad\le \sqrt{2}\kappa^{\frac{3}{4}}C_0\left(\|\hat{K} - K\|_2 + \|\hat{k}_x - k_x\| + |f_x|\right)^{1/4} \\ &\qquad + \sqrt{\kappa}C_0\left(\|\hat{K} - K\|_2 + \|\hat{k}_x - k_x\| + |f_x|\right)^{1/2},\end{aligned}$$

where $f_x = \hat{k}(x,x) - k(x,x)$.

Proof. Use the setup of Section 2.2 of Cortes et al. (2010). In particular, we will use $\|w\| \le \sqrt{\kappa}C_0$ and their (16-17):

$$\begin{aligned}\Phi(x_i) &= K_x^{1/2}e_i \\ \|\hat{w} - w\|^2 &\le 2C_0^2\sqrt{\kappa}\|\hat{K}_x^{1/2} - K_x^{1/2}\|,\end{aligned}$$

where $K_x = \begin{bmatrix} K & k_x \\ k_x^\mathsf{T} & k(x,x)\end{bmatrix}$ and e_i the ith standard basis.

Further, Lemma 1 of Cortes et al. (2010) says that $\|\hat{K}_x^{1/2} - K_x^{1/2}\|_2 \le \|\hat{K}_x - K_x\|_2^{1/2}$. Let $f_x := \hat{k}(x,x) - k(x,x)$; Then, by Weyl's inequality for singular values,

$$\left\|\begin{bmatrix}\hat{K} - K & \hat{k}_x - k_x \\ \hat{k}_x^\mathsf{T} - k_x^\mathsf{T} & f_x\end{bmatrix}\right\|_2 \le \|\hat{K} - K\|_2 + \|\hat{k}_x - k_x\| + |f_x|\,.$$

[5] We again assume there is no bias term for simplicity; adding a constant feature again changes the analysis only in that it makes the κ of Proposition 10 2 instead of 1.

Thus

$$\begin{aligned}
&|\hat{h}(x) - h(x)| \\
&\quad = \left|(\hat{w} - w)^\mathsf{T} \hat{\Phi}(x) + w^\mathsf{T} (\hat{\Phi}(x) - \Phi(x))\right| \\
&\quad \leq \|\hat{w} - w\| \|\hat{\Phi}(x)\| + \|w\| \|\hat{\Phi}(x) - \Phi(x)\| \\
&\quad \leq \sqrt{2} \kappa^{\frac{1}{4}} C_0 \|\hat{K}_x^{1/2} - K_x^{1/2}\|_2^{1/2} \sqrt{\kappa} \\
&\qquad + \sqrt{\kappa} C_0 \|(\hat{K}_x^{1/2} - K_x^{1/2}) e_{n+1}\| \\
&\quad \leq \sqrt{2} \kappa^{\frac{3}{4}} C_0 \|\hat{K}_x - K_x\|_2^{1/4} \\
&\qquad + \sqrt{\kappa} C_0 \|\hat{K}_x - K_x\|^{1/2} \\
&\quad \leq \sqrt{2} \kappa^{\frac{3}{4}} C_0 \left(\|\hat{K} - K\|_2 + \|\hat{k}_x - k_x\| + |f_x|\right)^{1/4} \\
&\qquad + \sqrt{\kappa} C_0 \left(\|\hat{K} - K\|_2 + \|\hat{k}_x - k_x\| + |f_x|\right)^{1/2}
\end{aligned}$$

as claimed. □

Suppose that $\sup|k(x,y) - s(x,y)| \leq \varepsilon$. Then, as in the last section, $\|\hat{k}_x - k_x\| \leq \sqrt{n}\varepsilon$ and $\|\hat{K} - K\|_2 \leq n\varepsilon$. Then, letting γ be 0 for $\tilde{z}$ and 1 for $\breve{z}$, Proposition 10 gives

$$\begin{aligned}
|\hat{h}(x) - h(x)| \leq{}& \sqrt{2} C_0 \left(n + \sqrt{n} + \gamma\right)^{1/4} \varepsilon^{1/4} \\
&+ C_0 \left(n + \sqrt{n} + \gamma\right)^{1/2} \varepsilon^{1/2}.
\end{aligned}$$

Then $|\hat{h}(x) - h(x)| \geq u$ only if

$$\varepsilon \leq \frac{2C_0^2 + 4C_0 u + u^2 - 2(C_0 + u)\sqrt{C_0(C_0 + 2u)}}{C_0^2(n + \sqrt{n} + \gamma)}.$$

This bound has the unfortunate property of requiring the approximation to be *more* accurate as the training set size increases, and thus can prove only a very loose upper bound on the number of features needed to achieve a given approximation accuracy, due to the looseness of Proposition 10. Analyses of generalization error in the induced RKHS, such as Rahimi and Recht (2008a) and Yang et al. (2012), are more useful in this case.

3.3 MAXIMUM MEAN DISCREPANCY

Another area of application for random Fourier embeddings is to the mean embedding of distributions, which uses some kernel k to represent a probability distribution P in the RKHS induced by k as $\varphi(P) = \mathbb{E}_{x \sim P}\left[k(x, \cdot)\right]$. For samples $\{X_i\}_{i=1}^n \sim P$ and $\{Y_j\}_{j=1}^m \sim Q$, we can estimate the inner product in the embedding space, the *mean map kernel* (MMK), by

$$\textsc{mmk}(X, Y) := \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m k(X_i, Y_j) \approx \left\langle \varphi(P), \varphi(Q) \right\rangle .$$

The distance $\|\varphi(P) - \varphi(Q)\|$ is known as the *maximum mean discrepancy* (MMD), which can be estimated with:

$$\begin{aligned}
&\|\varphi(P) - \varphi(Q)\|^2 \\
&= \langle \varphi(P), \varphi(P) \rangle + \langle \varphi(Q), \varphi(Q) \rangle - 2 \left\langle \varphi(P), \varphi(Q) \right\rangle .
\end{aligned}$$

$\textsc{mmk}(X, X)$ is a biased estimator, because of the $k(X_i, X_i)$ and $k(Y_i, Y_i)$ terms; removing them gives an unbiased estimator (Gretton et al. 2012). The MMK can be used in standard kernel methods to perform learning on probability distributions, such as when images are treated as sets of local patch descriptors (Muandet et al. 2012) or documents as sets of word descriptors (Yoshikawa et al. 2014). The MMD has strong applications to two-sample testing, where it serves as the statistic for testing the hypothesis that X and Y are sampled from the same distribution (Gretton et al. 2012); this has applications in, for example, comparing microarray data from different experimental situations or in matching attributes when merging databases.

The MMK estimate can clearly be approximated with an explicit embedding: if $k(x, y) \approx z(x)^\mathsf{T} z(y)$,

$$\begin{aligned}
\textsc{mmk}_z(X, Y) &= \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m z(X_i)^\mathsf{T} z(Y_j) \\
&= \left(\frac{1}{n} \sum_{i=1}^n z(X_i) \right)^\mathsf{T} \left(\frac{1}{m} \sum_{j=1}^m z(Y_j) \right) \\
&= \bar{z}(X)^\mathsf{T} \bar{z}(Y).
\end{aligned}$$

Thus the biased estimator of $\textsc{mmk}(X, X)$ is just $\|\bar{z}(X)\|^2$; the unbiased estimator is

$$\frac{n^2}{n^2 - n} \left(\|\bar{z}(X)\|^2 - \frac{1}{n^2} \sum_{i=1}^n \|z(X_i)\|^2 \right)$$

When $z(x)^\mathsf{T} z(x) = 1$, as with $\tilde{z}$, this simplifies to $\frac{n}{n-1} \|\bar{z}(X)\|^2 - \frac{1}{n-1}$. When that is not necessarily true, as with $\breve{z}$, that simplification holds only in expectation.

This has been noticed a few times in the literature, e.g. by Li and Tsang (2011). Gretton et al. (2012) gives different linear-time test statistics based on subsampling the sum over pairs; this version avoids reducing the amount of data used in favor of approximating the kernel. Additionally, when using the MMK in a kernel method this approximation allows the use of linear solvers, whereas the other linear approximations must still perform some pairwise computation. Zhao and Meng (2014) compare the empirical performance of an approximation equivalent to $\breve{z}$ against other linear-time approximations for two-sample testing. They find it is slower than the MMD-linear approximation but far more accurate, while being more accurate and comparable in speed to a block-based B-test (Zaremba et al. 2013).

Zhao and Meng (2014) also state a simple uniform error bound on the quality of this approximation. Specifically, since we can write $|\textsc{mmk}_z(X,Y) - \textsc{mmk}(X,Y)|$ as the mean of $|f(X_i,Y_j)|$, uniform error bounds on f apply directly to $\textsc{mmk}_z$, including to the unbiased version of $\textsc{mmk}_z(X,X)$. Moreover, since $\textsc{mmd}^2(X,Y) = \textsc{mmk}(X,X) + \textsc{mmk}(Y,Y) - 2\textsc{mmk}(X,Y)$, its error is at most 4 times $\|f\|_\infty$. The advantage of this bound is that it applies uniformly to all sample sets on the input space $\mathcal{X}$, which is useful when we use $\textsc{mmk}$ for a kernel method.

For a single two-sample test, however, we can get a tighter bound. Consider X and Y fixed for now. Note that $\mathbb{E}\textsc{mmk}_z(X,Y) = \textsc{mmk}(X,Y)$, by linearity of expectation. The variance of $\textsc{mmk}_z(X,Y)$ is exactly

$$\frac{1}{n^2m^2}\sum_{i,j}\sum_{i',j'}\operatorname{Cov}\left(s(X_i,Y_j), s(X_{i'},Y_{j'})\right), \qquad (12)$$

which can be evaluated using the formulas of Section 2.1 and so, viewed only as a function of D, is $O(1/D)$. Alternatively, we can use a bounded difference approach: viewing $\textsc{mmk}_{\tilde{z}}(X,Y)$ as a function of the ω_is, changing ω_i to $\hat{\omega}_i$ changes the $\textsc{mmk}$ estimate by

$$\left|\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{2}{D}\left(\cos(\hat{\omega}_i^\mathsf{T}(X_i-Y_j)) - \cos(\omega_i^\mathsf{T}(X_i-Y_j))\right)\right|,$$

which is at most $4/D$. The bound for $\breve{z}$ is in fact the same here. Thus McDiarmid's inequality tells us that for fixed sets X and Y and either z,

$$\Pr\left(|\textsc{mmk}_z(X,Y) - \textsc{mmk}(X,Y)|\right) \le 2\exp\left(-\tfrac{1}{8}D\varepsilon^2\right).$$

Thus $\mathbb{E}\,|\textsc{mmk}_z(X,Y) - \textsc{mmk}(X,Y)| \le 2\sqrt{2\pi/D}$. Similarly, $\textsc{mmd}_z$ can be changed by at most $16/D$, giving

$$\Pr\left(|\textsc{mmd}_z(X,Y) - \textsc{mmd}(X,Y)|\right) \le 2\exp\left(-\tfrac{1}{128}D\varepsilon^2\right)$$

and expected absolute error of at most $8\sqrt{2\pi/D}$.

Now, if we consider the distributions P and Q to be fixed but the sample sets random, Theorems 7 and 10 of Gretton et al. (2012) give exponential convergence bounds for the biased and unbiased population estimators of $\textsc{mmd}$, which can easily be combined with the above bounds. Note that this approach allows the domain $\mathcal{X}$ to be unbounded, unlike the other bound. One could extend this to a bound uniform over some smoothness class of distributions using the techniques of Section 2.2, though we do not do so here.

4 NUMERICAL EVALUATION

4.1 APPROXIMATION ON AN INTERVAL

We first conduct a detailed study of the approximations on the interval $\mathcal{X} = [-b,b]$. Specifically, we evenly spaced $1\,000$ points on $[-5,5]$ and approximated the kernel matrix using both embeddings at $D \in \{50, 100, 200, \dots, 900, 1\,000, 2\,000, \dots, 9\,000, 10\,000\}$, repeating each trial $1\,000$ times, estimating $\|f\|_\infty$ and $\|f\|_\mu$ at those points. We do not consider $d > 1$ here, because obtaining a reliable estimate of $\sup|f|$ becomes very computationally expensive even for $d = 2$.

Figure 3 shows the behavior of $\mathbb{E}\|f\|_\infty$ as b increases for various values of D. As expected, the $\tilde{z}$ embeddings have almost no error near 0. The error increases out to one or two bandwidths, after which the curve appears approximately linear in ℓ/σ, as predicted by Proposition 3.

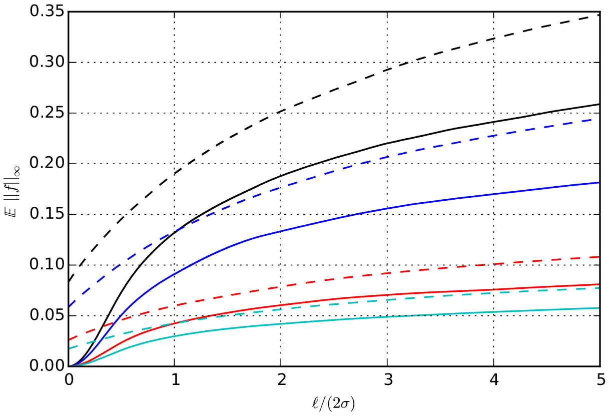

Figure 3: The maximum error within a given radius in $\mathbb{R}$, averaged over $1\,000$ evaluations. Solid lines represent $\tilde{z}$ and dashed lines $\breve{z}$; black is $D = 50$, blue is $D = 100$, red $D = 500$, and cyan $D = 1\,000$.

Figure 4 fixes $b = 3$ and shows the expected maximal error as a function of D. It also plots the expected error obtained by numerically integrating the bounds of Propositions 1 and 2 (using the minimum of 1 and the bound). We can see that all of the bounds are fairly loose, but that the first version of the bound in the propositions (with β_d, the exponent depending on d, and α_ε) is substantially tighter than the second version when $d = 1$.

The bounds on $\mathbb{E}\|f\|_\infty$ of Propositions 3 and 4 are unfortunately too loose to show on the same plot. However, one important property does hold. For a fixed $\mathcal{X}$, (8) predicts that $\mathbb{E}\|f\|_\infty = O(1/\sqrt{D})$. This holds empirically: performing linear regression of $\log\mathbb{E}\|\tilde{f}\|_\infty$ against $\log D$ yields a model of $\mathbb{E}\|\tilde{f}\|_\infty = e^c D^m$, with a 95% confidence interval for m of $[-0.502, -0.496]$; $\|\breve{f}\|_\infty$ gives $[-0.503, -0.497]$. The integrated bounds of Propositions 1 and 2 do not fit the scaling as a function of D nearly as well.

Figure 5 shows the empirical survival function of the max error for $D = 500$, along with the bounds of Propositions 1 and 2 and those of Propositions 5 and 6 using the empirical mean. The latter bounds are tighter than the former for low ε, especially for low D, but have a lower slope.

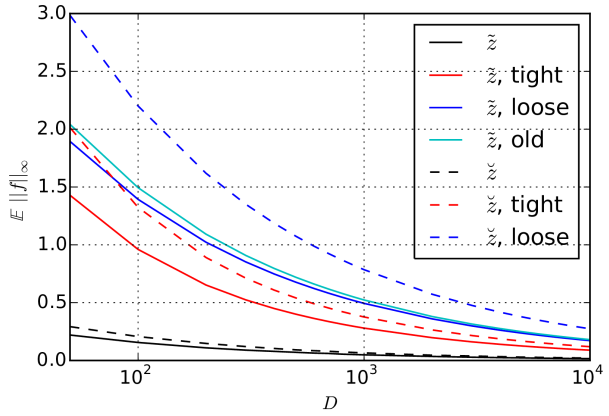

Figure 4: $\mathbb{E}\|f\|_\infty$ for the Gaussian kernel on $[-3, 3]$ with $\sigma = 1$, based on the mean of 1 000 evaluations and on numerical integration of the bounds from Propositions 1 and 2. ("Tight" refers to the bound with constants depending on d, and "loose" the second version; "old" is the version from Rahimi and Recht (2007).)

The mean of the mean squared error, on the other hand, exactly follows the expectation of Section 2.3 using μ as the uniform distribution on $\mathcal{X}^2$: in this case, $\mathbb{E}\|\tilde{f}\|_\mu \approx 0.66/D$, $\mathbb{E}\|\breve{f}\|_\mu \approx 0.83/D$. (This is natural, as the expectation is exact.) Convergence to that mean, however, is substantially faster than guaranteed by the McDiarmid bound of Propositions 7 and 8. We omit the plot due to space constraints.

4.2 MAXIMUM MEAN DISCREPANCY

We now turn to the problem of computing the MMD with a Fourier embedding. Specifically, we consider the problem of distinguishing the standard normal distribution $\mathcal{N}(0, I_p)$ from the two-dimensional mixture $0.95\mathcal{N}(0, I_2) + 0.05\mathcal{N}(0, \frac{1}{4}I_2)$. We take fixed sample sets X and Y each of size 1 000 and compute the biased MMD estimate with varying D for both $\tilde{z}$ and $\breve{z}$, we used a Gaussian kernel of bandwidth 1. The mean absolute errors of the resulting estimates are shown in Figure 6. $\tilde{z}$ performs mildly better than $\breve{z}$.

Again, the McDiarmid bound of Section 3.3 predicts that the mean absolute error decays as $O(1/\sqrt{D})$, but with too high a multiplicative constant; the 95% confidence interval for the exponent of D is $[-0.515, -0.468]$ for $\tilde{z}$ and $[-0.520, -0.486]$ for $\breve{z}$. We also know that the expected root mean squared error decays like $O(1/\sqrt{D})$ via (12).

5 DISCUSSION

We provide a novel investigation of the approximation error of the popular random Fourier features, tightening existing bounds and showing new ones, including an analytic bound on $\mathbb{E}\|f\|_\infty$ and exponential concentration about its mean, as well as an exact form for $\mathbb{E}\|f\|_\mu$ and exponential concentration in that case as well. We also extend previous results on the change in learned models due to kernel approximation. We verify some aspects of these bounds empirically for the Gaussian kernel. We also point out that, of the two embeddings provided by Rahimi and Recht (2007), the $\tilde{z}$ embedding (with half as many sampled frequencies, but no additional noise due to phase shifts) is superior in the most common case of the Gaussian kernel.

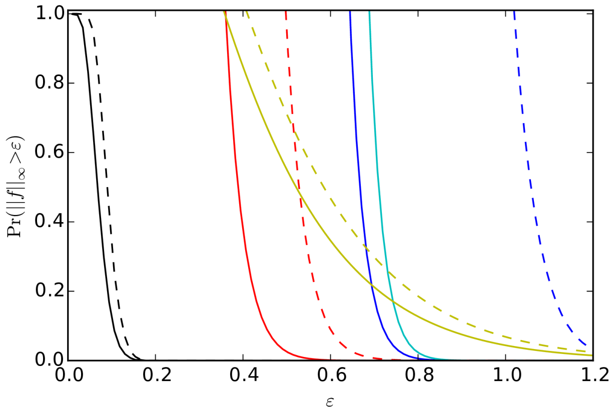

Figure 5: $\Pr\left(\mathbb{E}\|f\|_\infty > \varepsilon\right)$ for the Gaussian kernel on $[-3, 3]$ with $\sigma = 1$ and $D = 500$, based on 1 000 evaluations (black), numerical integration of the bounds from Propositions 1 and 2 (same colors as Figure 4), and the bounds of Propositions 5 and 6 using the empirical mean (yellow).

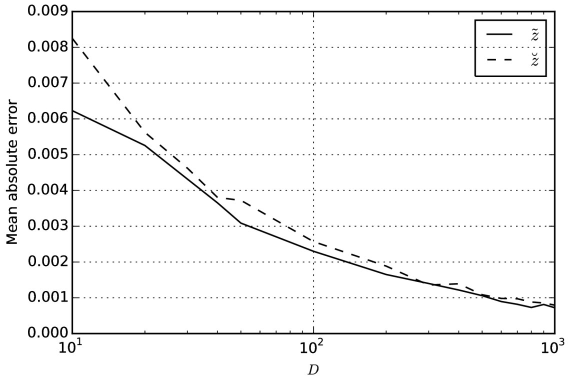

Figure 6: Mean absolute error of the biased estimator for $\text{MMD}(X, Y)$, based on 100 evaluations.

Acknowledgments

This work was funded in part by DARPA grant FA87501220324. DJS is also supported by a Sandia Campus Executive Program fellowship.

References

Bernstein, Sergei (1924). “On a modification of Chebyshevs inequality and of the error formula of Laplace”. Russian. In: *Ann. Sci. Inst. Savantes Ukraine, Sect. Math.* 1, pp. 38–49.

Bochner, Salomon (1959). *Lectures on Fourier integrals*. Princeton University Press.

Boucheron, Stéphane, Gábor Lugosi, and Pascal Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, UK: Oxford University Press.

Bousquet, Olivier (2002). “A Bennett concentration inequality and its application to suprema of empirical processes”. In: *Comptes Rendus Mathematique* 334, pp. 495–500.

Bousquet, Olivier and André Elisseeff (2001). “Algorithmic Stability and Generalization Performance”. In: *Advances in Neural Information Processing Systems*, pp. 196–202.

Cheng, Steve (2013). *Differentiation under the integral sign*. Version 16. URL: http://planetmath.org/differentiationundertheintegralsign.

Cortes, Corinna, M Mohri, and A Talwalkar (2010). “On the impact of kernel approximation on learning accuracy”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 113–120.

Cucker, Felipe and Steve Smale (2001). “On the mathematical foundations of learning”. In: *Bulletin of the American Mathematical Society* 39.1, pp. 1–49.

Dai, Bo et al. (2014). “Scalable Kernel Methods via Doubly Stochastic Gradients”. In: *Advances in Neural Information Processing Systems*, pp. 3041–3049.

Dudley, Richard M (1967). “The sizes of compact subsets of Hilbert space and continuity of Gaussian processes”. In: *Journal of Functional Analysis* 1.3, pp. 290–330.

Gretton, Arthur, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alex J Smola (2012). “A Kernel Two-Sample Test”. In: *The Journal of Machine Learning Research* 13.

Hoeffding, Wassily (1963). “Probability inequalities for sums of bounded random variables”. In: *Journal of the American Statistical Association* 58.301, pp. 13–30.

Joachims, Thorsten (2006). “Training linear SVMs in linear time”. In: *ACM SIGKDD international conference on Knowledge Discovery and Data mining*.

Li, Shukai and Ivor W Tsang (2011). “Learning to Locate Relative Outliers”. In: *Asian Conference on Machine Learning*. Vol. 20. JMLR: Workshop and Conference Proceedings, pp. 47–62.

McDiarmid, Colin (1989). “On the method of bounded differences”. In: *Surveys in combinatorics* 141.1, pp. 148–188.

Muandet, Krikamol, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf (2012). “Learning from distributions via support measure machines”. In: *Advances in Neural Information Processing Systems*.

Pedregosa, F. et al. (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Raff, Edward (2011-15). *JSAT: Java Statistical Analysis Tool*. https://code.google.com/p/java-statistical-analysis-tool/.

Rahimi, Ali and Benjamin Recht (2007). “Random Features for Large-Scale Kernel Machines”. In: *Advances in Neural Information Processing Systems*. MIT Press.

– (2008a). “Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning”. In: *Advances in Neural Information Processing Systems*. MIT Press, pp. 1313–1320.

– (2008b). “Uniform approximation of functions with random bases”. In: *46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 555–561.

Saunders, C., A. Gammerman, and V. Vovk (1998). “Ridge Regression Learning Algorithm in Dual Variables”. In: *Proceedings of the 15th International Conference on Machine Learning*, pp. 515–521.

Sonnenburg, Sören et al. (2010). “The SHOGUN Machine Learning Toolbox”. In: *Journal of Machine Learning Research* 11, pp. 1799–1802.

Yang, Tianbao, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou (2012). “Nyström Method vs Random Fourier Features: A Theoretical and Empirical Comparison”. In: *Advances in Neural Information Processing Systems*. MIT Press.

Yoshikawa, Yuya, Tomoharu Iwata, and Hiroshi Sawada (2014). “Latent Support Measure Machines for Bag-of-Words Data Classification”. In: *Advances in Neural Information Processing Systems*, pp. 1961–1969.

Zaremba, Wojciech, Arthur Gretton, and Matthew Blaschko (2013). “B-tests: Low Variance Kernel Two-Sample Tests”. In: *Advances in Neural Information Processing Systems*.

Zhao, Ji and Deyu Meng (2014). “FastMMD: Ensemble of Circular Discrepancy for Efficient Two-Sample Test”. In: arXiv: 1405.2664.