On the Error of Random Fourier Features

Dougal J. Sutherland and Jeff Schneider

{dsutherl,schneide}@cs.cmu.edu

Auton Lab

Carnegie Mellon University

Random Fourier features

Random Fourier features (Rahimi and Recht, 2007) scale shift-invariant kernels to large numbers of inputs by using linear models on $z(x)$, where $z : \mathbb{R}^d \to \mathbb{R}^D$ has $k(x,y) \approx z(x)^\mathsf{T} z(y)$.

Let $\Delta := x - y$, and $k(x,y) = k(\Delta)$, $k(0) = 1$ be a continuous PSD kernel. Its Fourier transform $P(\omega)$ is a probability measure (Bochner's theorem).

One embedding is:

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

$\phi(x)^\mathsf{T}\phi(y)$ is an average of $D/2$ terms $\cos(\omega_i^\mathsf{T}\Delta)$; note $\mathbb{E}\cos(\omega^\mathsf{T}\Delta) = \Re \int e^{\omega^\mathsf{T}\Delta\sqrt{-1}} \, \mathrm{d}P(\omega) = k(\Delta)$.

Another has more samples from $P(\omega)$, but with additional non-shift-invariant noise:

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

$\psi(x)^\mathsf{T}\psi(y)$ is the mean of D terms of the form $\cos(\omega_i^\mathsf{T}\Delta) + \cos(\omega_i^\mathsf{T}(x+y) + 2b_i)$.

Our contribution

We show the ϕ embedding has lower variance than ψ for the Gaussian kernel, and improve the theoretical understanding of both embeddings' errors.

Prevalence of the embeddings

- The original publication discussed both ϕ and ψ. Online revisions only mention ψ (but give a bound only for ϕ). Later work used only ψ.
- Of the first 100 citations on Google Scholar, 15 used ψ, 14 used ϕ, 28 did not specify.
- All three library implementations we found (scikit-learn, Shogun, and JSAT) use ψ.

Variance

Using trig identities, we can show that

$$\operatorname{Var} \phi(x)^\mathsf{T}\phi(y) = \frac{1}{D}\left[1 + k(2\Delta) - 2k(\Delta)^2\right]$$

$$\operatorname{Var} \psi(x)^\mathsf{T}\psi(y) = \frac{1}{D}\left[1 + \tfrac{1}{2}k(2\Delta) - k(\Delta)^2\right].$$

So ϕ is lower-variance when

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

ϕ is better for Gaussian kernels

For $k(\Delta) := \exp\left(-\|\Delta\|^2/(2\sigma^2)\right)$,

$$\operatorname{Var}\cos(\omega) = \tfrac{1}{2}\left(1 - \exp\left(-\|\Delta\|^2/\sigma^2\right)\right) \le \tfrac{1}{2}.$$

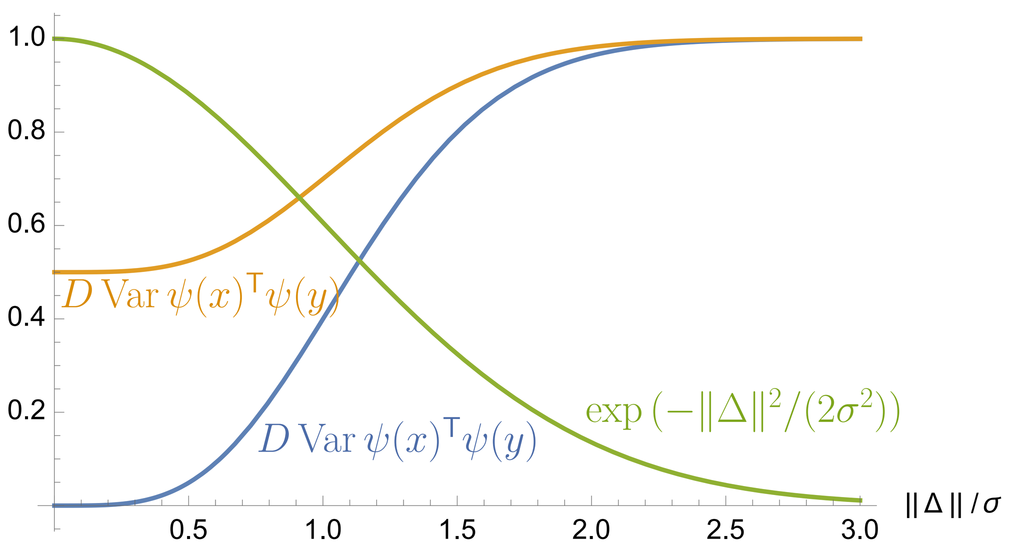

Figure 1: The variance per dimension for the Gaussian kernel. The difference in variance is higher for larger kernel values.

Improved uniform convergence

We can tighten the bound for ϕ and show one for ψ.

- Let ℓ be the diameter of the domain $\mathcal{X} \subset \mathbb{R}^d$.
- Let $\sigma_p^2 := \mathbb{E}\|\omega\|^2$, $\sigma_w^2 := \sup_\Delta \left[2 \operatorname{Var}\cos(\omega^\mathsf{T}\Delta)\right]$.

Define $f_\phi(x,y) := \phi(x)^\mathsf{T}\phi(y) - k(x,y)$ to be the error for ϕ, and let $\alpha_\varepsilon := \min\left(1, \frac{1}{2}\sigma_w^2 + \frac{1}{3}\varepsilon\right)$; then

$$\Pr\left(\|f_\phi\|_\infty \ge \varepsilon\right) \le \beta_d \left(\frac{\sigma_p \ell}{\varepsilon}\right)^{\frac{2}{1+\frac{2}{d}}} \exp\left(-\frac{D\varepsilon^2}{8(d+2)\alpha_\varepsilon}\right).$$

For ψ, define $f_\psi(x,y) := \psi(x)^\mathsf{T}\psi(y) - k(x,y)$ as well as $\alpha'_\varepsilon := \min\left(1, \frac{1}{8}(1+\sigma_w^2) + \frac{1}{6}\varepsilon\right)$; then

$$\Pr\left(\|f_\psi\|_\infty \ge \varepsilon\right) \le \beta'_d \left(\frac{\sigma_p \ell}{\varepsilon}\right)^{\frac{2}{1+\frac{1}{d}}} \exp\left(-\frac{D\varepsilon^2}{32(d+2)\alpha'_\varepsilon}\right).$$

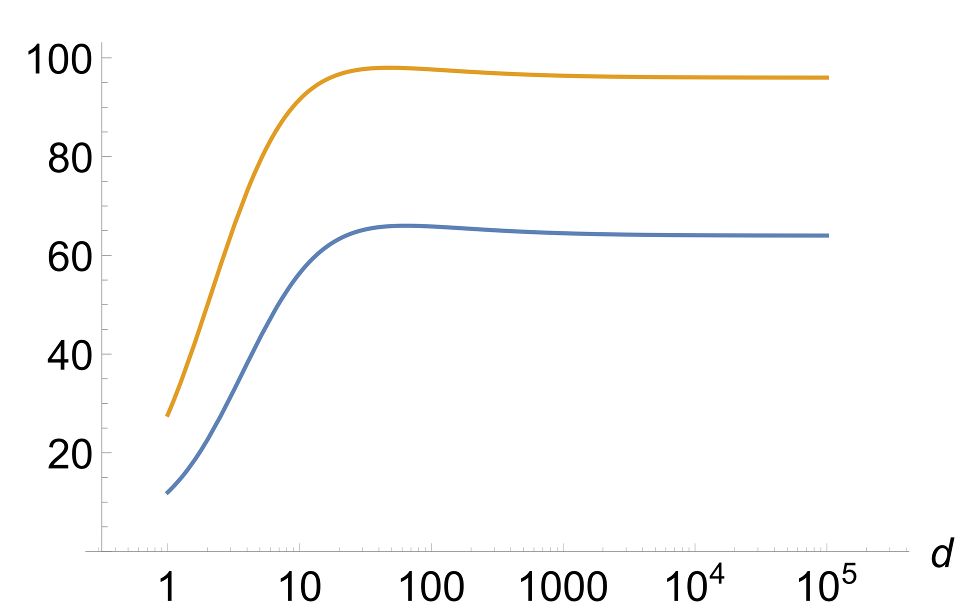

Figure 2: The coefficients β_d (blue, for ϕ) and β'_d (orange, for ψ).

The bound for ϕ is always tighter than that for ψ.

Expected max error

- Suppose $k(\Delta)$ is L-Lipschitz.
- Let $\gamma \approx 0.964$, $0.8 < \gamma' < 1.55$ depending on $\mathcal{X}$.

$$\mathbb{E}\|f_\phi\|_\infty \le \frac{24\gamma\sqrt{d}\ell}{\sqrt{D}}\left(L + \mathbb{E}\max_{i=1,\dots,\frac{D}{2}}\|\omega_i\|\right),$$

$$\mathbb{E}\|f_\psi\|_\infty \le \frac{48\gamma'\sqrt{d}\ell}{\sqrt{D}}\left(L + \mathbb{E}\max_{i=1,\dots,D}\|\omega_i\|\right)$$

using Dudley's entropy integral.

Concentration

$$\Pr\left(\|f_\phi\|_\infty \ge \mathbb{E}\|f_\phi\|_\infty + \varepsilon\right) \le 2\exp\left(-\frac{D\varepsilon^2}{D\,\mathbb{E}\|f_\phi\|_\infty + \frac{1}{2}\sigma_w^2 + \frac{1}{6}D\varepsilon}\right),$$

$$\Pr\left(\|f_\psi\|_\infty \ge \mathbb{E}\|f_\psi\|_\infty + \varepsilon\right) \le 2\exp\left(-\frac{D\varepsilon^2}{\frac{4}{9}D\,\mathbb{E}\|f_\psi\|_\infty + \frac{1}{81}(1+\sigma_w^2) + \frac{2}{27}D\varepsilon}\right)$$

via Bousquet's inequality. f_ψ concentrates more tightly, but its mean is higher, both in the bound and empirically.

Numerical results with $d = 1$

Gaussian kernel, $\sigma = 1$. ϕ has solid lines, ψ dashed.

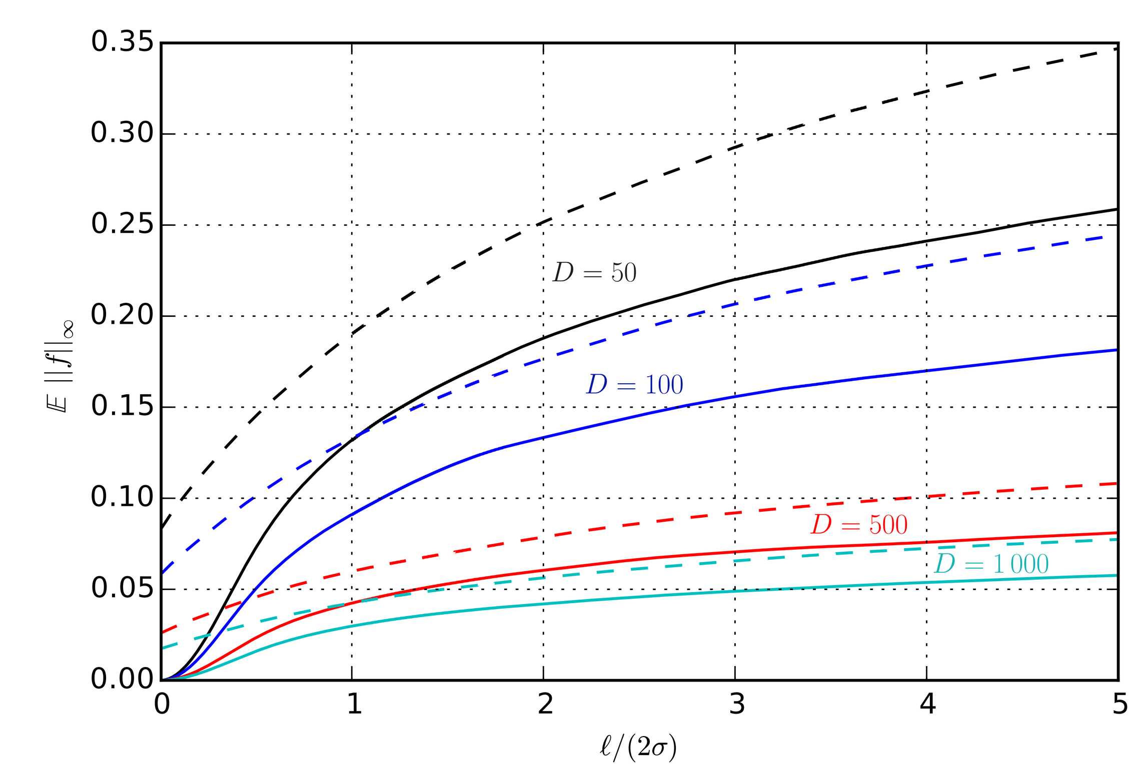

Figure 3: Average max error within a given radius.

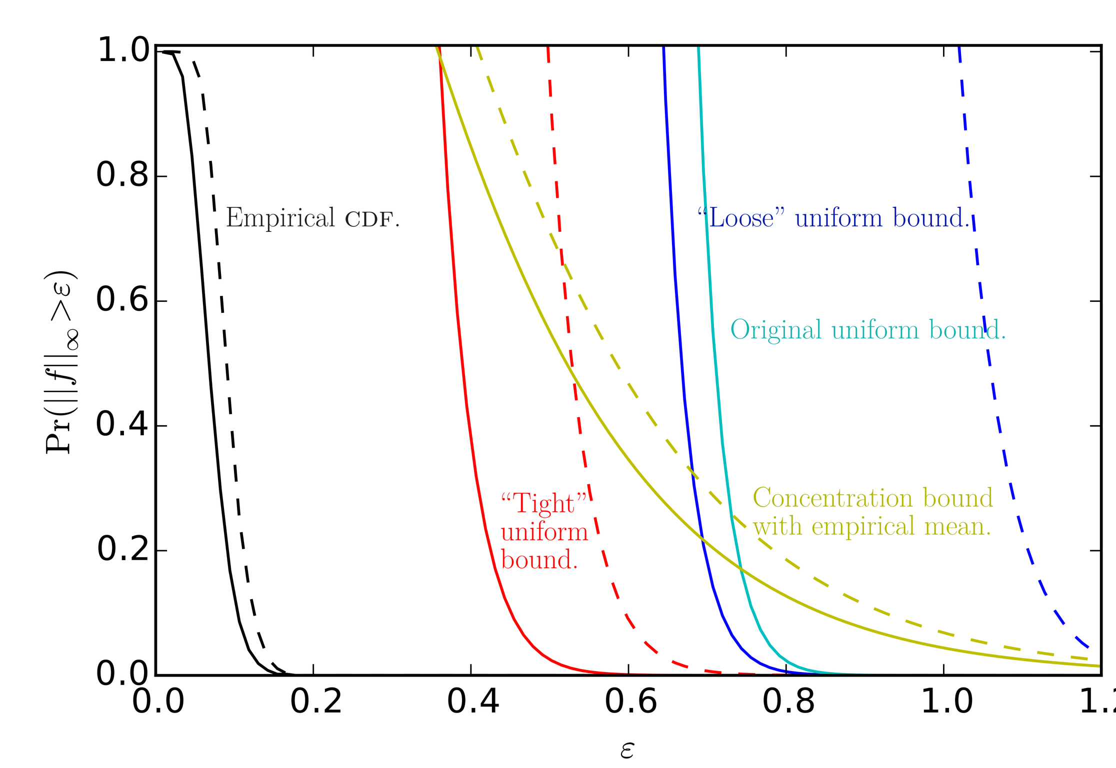

Figure 4: $\Pr\left(\|f\|_\infty > \varepsilon\right)$ on $[-3, 3]$ with $D = 500$.

Further results

The paper also has:

- Exact expectations and contentration bounds of squared L_2 error, for any measure.
- Bounds on changes in the outputs of ridge regression, SVM, and maximum mean discrepancy tests due to the features.
- More experiments.

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