Understanding Random Kitchen Sinks

Edwin V. Bonilla

August 13, 2015

In this paper we make sense of the Random Kitchen Sinks (RKS) method (add ref) , which approximates a kernel using a finite set of random features. FIX

1 Approximations for Stationary Kernels

The main starting point is Bochner's theorem regarding stationary processes, which in turn yields Wiener-Khintchine's theorem on the Fourier duality of the covariance function of the process and its spectral density:

$$\kappa(\boldsymbol{\tau}) = \int f(\mathbf{w}) e^{2\pi i \mathbf{w}^T \boldsymbol{\tau}} d\mathbf{w}, \tag{1}$$

$$f(\mathbf{w}) = \int \kappa(\boldsymbol{\tau}) e^{-2\pi i \mathbf{w}^T \boldsymbol{\tau}} d\boldsymbol{\tau}. \tag{2}$$

Ali Rahimi's main insight is that we can approximate the above kernel by explicitly constructing "suitable" features and (Monte Carlo) averaging over samples from $f(\mathbf{w})$:

$$\kappa(\mathbf{x}, \mathbf{y}) = \kappa(\mathbf{x} - \mathbf{y}) = \kappa(\boldsymbol{\tau}) = \mathbb{E}_{\mathbf{w}}[\boldsymbol{\phi}_{\mathbf{w}}(\mathbf{x})^T \boldsymbol{\phi}_{\mathbf{w}}(\mathbf{y})] \tag{3}$$

$$\approx \frac{1}{D} \sum_{i=1}^{D} \phi_{\mathbf{w}_i}(\mathbf{x}) \phi_{\mathbf{w}_i}(\mathbf{y}) \tag{4}$$

1.1 Features

As the kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa(\mathbf{x} - \mathbf{y}) = \kappa(\boldsymbol{\tau})$ is real, and the density $f(\mathbf{w})$ is real, then we only care about the real part:

$$\Re\{e^{2\pi i \mathbf{w}^T \boldsymbol{\tau}}\} = \cos(2\pi \mathbf{w}^T \boldsymbol{\tau}) = \cos(2\pi \mathbf{w}^T (\mathbf{x} - \mathbf{y})) \tag{5}$$

If we use features of the form:

$$\boldsymbol{\phi}_{\mathbf{w}}(\mathbf{x}) = [\cos(2\pi \mathbf{w}^T \mathbf{x}), \sin(2\pi \mathbf{w}^T \mathbf{x})]^T, \tag{6}$$

Then we have that:

$$\boldsymbol{\phi}_{\mathbf{w}}(\mathbf{x})^T \boldsymbol{\phi}_{\mathbf{w}}(\mathbf{y}) = \cos(2\pi \mathbf{w}^T \mathbf{x}) \cos(2\pi \mathbf{w}^T \mathbf{y}) + \sin(2\pi \mathbf{w}^T \mathbf{x}) \sin(2\pi \mathbf{w}^T \mathbf{y}) \tag{7}$$

$$= \cos(2\pi \mathbf{w}^T (\mathbf{x} - \mathbf{y})), \tag{8}$$

where we have obtained the features of the type required by Equation (5). Now, we can define our vector of features as:

$$\boldsymbol{\phi}_{\mathbf{w}}(\mathbf{x}) = \frac{1}{\sqrt{D}} [\cos(2\pi \mathbf{w}_1^T \mathbf{x}), \ldots, \cos(2\pi \mathbf{w}_D^T \mathbf{x}), \ldots \sin(2\pi \mathbf{w}_1^T \mathbf{x}), \ldots \sin(2\pi \mathbf{w}_D^T \mathbf{x})]^T \tag{9}$$

so that when computing the dot product $\boldsymbol{\phi}_{\mathbf{w}}(\mathbf{x})^T \boldsymbol{\phi}_{\mathbf{w}}(\mathbf{y})$ we obtain the approximation in Equation (4).

1.2 Density for the Weights

A question remains on what density should the weights be sampled from. This, of course, depends on the kernel. For the isotropic squared exponential covariance function we have:

$$\kappa(\boldsymbol{\tau}) = \exp(-\frac{1}{2\ell^2}\|\boldsymbol{\tau}\|^2), \tag{10}$$

where $\boldsymbol{\tau}$ lies in the original input dimensional space $\Re^d$. Having this, it is a matter of computing the Fourier transform in Equation (17). We will derive the density for the weights corresponding to this covariance function in 1-dimension as the Fourier transform of decomposable functions is simply the product of the individual Fourier transforms.

The Fourier transform of $\kappa(\tau) = \exp(-\frac{1}{2\ell^2}\tau^2)$ is given by:

$$\begin{aligned}
f(w) &= \sqrt{2\pi\ell^2}\exp(-2\pi^2\ell^2 w^2) && (11)\\
&= \sqrt{2\pi\ell^2}\exp\left(-\frac{1}{2}w^2\frac{1}{(2\pi\ell)^{-2}}\right) && (12)\\
&= \sqrt{2\pi\ell^2}\mathcal{N}(w;0,(2\pi\ell)^{-2})\frac{\sqrt{2\pi}}{2\pi\ell} \quad \text{Made the exponential a normalized Gaussian} && (13)\\
&= \frac{1}{2\pi\ell}\mathcal{N}(w;0,1) && (14)\\
&= \sigma_w\mathcal{N}(w;0,1), && (15)
\end{aligned}$$

where $\sigma_w^2 = 1/(2\pi\ell)^2$ is the variance of the Gaussian from which the weights are to be sampled. This is the optimal setting when the length-scale of the GP is known. For D new features and d input dimensions, we simply sample $D \times d$ standard Gaussian variables and multiply them by σ_w, which becomes a parameter of the features, as we may want to optimize this when the true length-scale of the process is unknown.

1.3 Other Conventions for the Fourier Transforms

Ali Rahimi, in fact, follows a different convention for the Fourier and Inverse Fourier Transforms:

$$\kappa(\boldsymbol{\tau}) = \int f(\mathbf{w})e^{i\mathbf{w}^T\boldsymbol{\tau}}d\mathbf{w}, \tag{16}$$

$$f(\mathbf{w}) = \frac{1}{(2\pi)^d}\int \kappa(\boldsymbol{\tau})e^{-i\mathbf{w}^T\boldsymbol{\tau}}d\boldsymbol{\tau}. \tag{17}$$

Following a similar derivation as above we find that the features are given by:

$$\boldsymbol{\phi}_{\mathbf{w}}(\mathbf{x}) = \frac{1}{\sqrt{D}}[\cos(\mathbf{w}_1^T\mathbf{x}),\ldots,\cos(\mathbf{w}_D^T\mathbf{x}),\ldots\sin(\mathbf{w}_1^T\mathbf{x}),\ldots\sin(\mathbf{w}_D^T\mathbf{x})]^T \tag{18}$$

and the density of the weights corresponding to the squared exponential covariance function with length scale ℓ is given by:

$$f(w) = \frac{1}{\ell}\mathcal{N}(w;0,1), \tag{19}$$

in other words, $\sigma_w = \ell^{-1}$.

2 Experiments

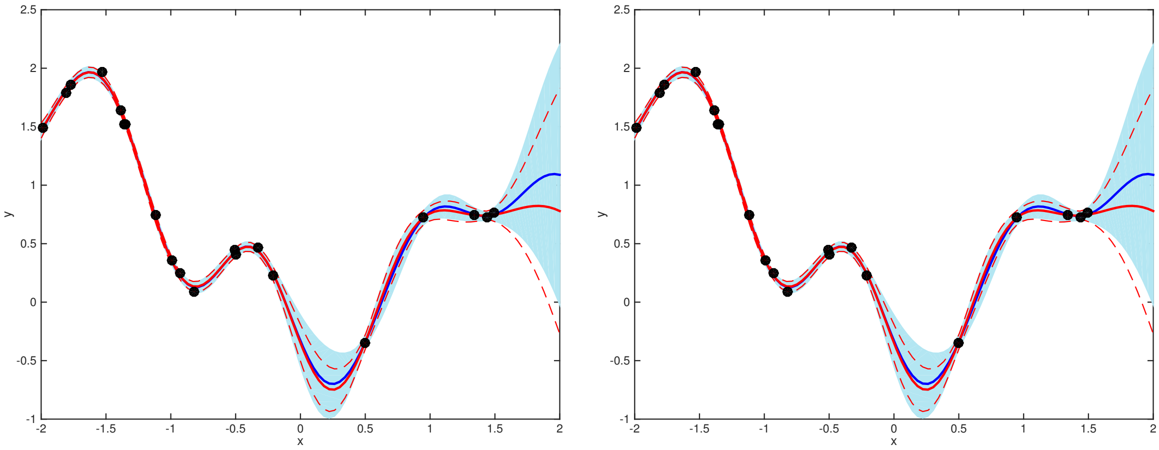

Figure 1: Experiments on 1d input and 1000 features (bases). Left: using the "traditional FT convention". Right: Using Ali Rahimi's convention. The red curves are the mean and two standard errors of the predictive distribution by a GP. The blue line and colored area are the same measures using RKS.