Random Fourier features for sampling from Gaussian processes

Matthew W. Hoffman

March 28, 2014

1 Gaussian processes

A Gaussian process (GP) is a stochastic process over an index set $\mathcal{X}$ that is fully specified by its mean and kernel functions, $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ respectively. This process is often used as a prior distribution over functions $f : \mathcal{X} \to \mathbb{R}$ which we will denote as $f \sim \mathrm{GP}(m, k)$. In this setting, adopting a GP prior means that given any finite collection of inputs $\mathbf{x}_{1:t}$ the outputs will be jointly Gaussian,

$$\mathbf{f}(\mathbf{x}_{1:t}) \sim \mathcal{N}(\mathbf{m}(\mathbf{x}_{1:t}), \mathbf{K}(\mathbf{x}_{1:t}, \mathbf{x}_{1:t})),$$

here $\mathbf{K}(\cdot, \cdot)$ denotes the pairwise kernel matrix; $\mathbf{f}(\cdot)$ and $\mathbf{m}(\cdot)$ respectively denote the latent and mean functions evaluated element-wise. For simplicity we will assume a zero mean prior, i.e. where $m(\mathbf{x}) = 0$.

Given a GP prior over function values we can also condition on observed input/output pairs in order to make predictions at test inputs. In order to do so we will also assume observations of the function at any point $\mathbf{x}_t$ are subject to Gaussian noise, i.e. $y_t \sim \mathcal{N}(f(\mathbf{x}_t), \sigma^2)$. After making t observations $\mathcal{D}_t = \{\mathbf{x}_{1:t}, \mathbf{y}_{1:t}\}$ we can write the joint distribution of the data and some test evaluation at $\mathbf{x}$ as

$$\begin{bmatrix} \mathbf{y}_{1:t} \\ f(\mathbf{x}) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_t + \sigma^2\mathbf{I} & \mathbf{k}_t(\mathbf{x}) \\ \mathbf{k}_t(\mathbf{x})^\mathsf{T} & k(\mathbf{x}, \mathbf{x}) \end{bmatrix}\right).$$

where $\mathbf{K}_t = \mathbf{K}(\mathbf{x}_{1:t}, \mathbf{x}_{1:t})$ and $\mathbf{k}_t(\mathbf{x}) = \mathbf{k}(\mathbf{x}_{1:t}, \mathbf{x})$. By a simple application of the matrix inversion lemma we can see that $f(\mathbf{x})|\mathcal{D}_t$ for the test evaluation is Gaussian with mean and variance given by

$$\mu_t(\mathbf{x}) = \mathbf{k}_t(\mathbf{x})^\mathsf{T}(\mathbf{K}_t + \sigma^2\mathbf{I})^{-1}\mathbf{y}_{1:t}, \tag{1}$$

$$\sigma_t^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_t(\mathbf{x})^\mathsf{T}(\mathbf{K}_t + \sigma^2\mathbf{I})^{-1}\mathbf{k}_t(\mathbf{x}). \tag{2}$$

2 Random Fourier features

In this section we will show that a shift-invariant kernel can be written as the expected inner-product of features taken from its Fourier transform. This section will roughly follow that of Rahimi and Recht [2007]. The following theorem, due to Bochner [1959], allows us to directly relate k to its Fourier transform:

Theorem 1 (Bochner's theorem)**.** *A continuous, shift-invariant kernel is positive definite if and only if it is the Fourier transform of a non-negative measure.*

A result of this theorem is that if the associated measure has density $s(\mathbf{w})$, known as the *spectral density*, then this density and the kernel are Fourier duals which can be written as

$$k(\mathbf{x}, \mathbf{x}') = \int e^{-i\mathbf{w}^\mathsf{T}(\mathbf{x}-\mathbf{x}')} s(\mathbf{w})\, d\mathbf{w},$$
$$s(\mathbf{w}) = \frac{1}{(2\pi)^d} \int e^{i\mathbf{w}^\mathsf{T}\boldsymbol{\tau}} k(\boldsymbol{\tau}, \mathbf{0})\, d\boldsymbol{\tau}.$$

Further, we can treat this measure as a probability density $p(\mathbf{w}) = s(\mathbf{w})/\alpha$ where $\alpha = \int s(\mathbf{w})\, d\mathbf{w}$ is the normalizing constant. Consequently, the kernel can be written as

$$k(\mathbf{x}, \mathbf{x}') = \alpha\, \mathbb{E}_\mathbf{w}[e^{-i\mathbf{w}^\mathsf{T}(\mathbf{x}-\mathbf{x}')}]$$

and due to the symmetry of $p(\mathbf{w})$ [see Rasmussen and Williams, 2006] we can write the expectation as

$$= \alpha\, \mathbb{E}_\mathbf{w}[\tfrac{1}{2}(e^{-i\mathbf{w}^\mathsf{T}(\mathbf{x}-\mathbf{x}')} + e^{i\mathbf{w}^\mathsf{T}(\mathbf{x}-\mathbf{x}')})]$$
$$= \alpha\, \mathbb{E}_\mathbf{w}[\cos(\mathbf{w}^\mathsf{T}\mathbf{x} - \mathbf{w}^\mathsf{T}\mathbf{x}')]$$

We can then note that $\int_0^{2\pi} \cos(a + 2b)\, db = 0$ for any constant offset $a \in \mathbb{R}$. As a result, for $b \sim \mathcal{U}(0, 2\pi)$ we can write

$$= \alpha\, \mathbb{E}_\mathbf{w}[\cos(\mathbf{w}^\mathsf{T}\mathbf{x} - \mathbf{w}^\mathsf{T}\mathbf{x}') + \mathbb{E}_b[\cos(\mathbf{w}^\mathsf{T}\mathbf{x} + \mathbf{w}^\mathsf{T}\mathbf{x}' + 2b)]]$$
$$= \alpha\, \mathbb{E}_{\mathbf{w},b}[\cos(\mathbf{w}^\mathsf{T}\mathbf{x} + b - \mathbf{w}^\mathsf{T}\mathbf{x}' - b) + \cos(\mathbf{w}^\mathsf{T}\mathbf{x} + b + \mathbf{w}^\mathsf{T}\mathbf{x}' + b)]$$
$$= 2\alpha\, \mathbb{E}_{\mathbf{w},b}[\cos(\mathbf{w}^\mathsf{T}\mathbf{x} + b)\cos(\mathbf{w}^\mathsf{T}\mathbf{x}' + b)]$$

The last equality can be derived from the sum of angles formula, which leads to the identity: $2\cos(x)\cos(y) = \cos(x - y) + \cos(x + y)$. Finally, given N weights and phases sampled from $p(\mathbf{w}, b)$ the kernel can be approximated as

$$\approx \frac{2\alpha}{N} \sum_{i=1}^{N} \cos(\mathbf{w}_i^\mathsf{T}\mathbf{x} + b_i)\cos(\mathbf{w}_i^\mathsf{T}\mathbf{x}' + b_i).$$

Collecting these samples as $[\mathbf{W}]_i = \mathbf{w}_i^\mathsf{T}$ and $[\mathbf{b}]_i = b_i$ we can introduce the feature mapping

$$\boldsymbol{\phi}(\mathbf{x}) = \sqrt{2\alpha/N}\cos(\mathbf{W}\mathbf{x} + \mathbf{b})$$

and it necessarily follows that our kernel function can be approximated as the inner product of these features $k(\mathbf{x}, \mathbf{x}') \approx \boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}')$

2.1 Predicting and sampling with random Fourier features

Given a set of features $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^N$ as defined earlier, we can now consider a linear model $f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\boldsymbol{\theta}$ with a standard normal prior over the weight vector $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We will again consider t input/output observations $\mathcal{D}_t = \{\mathbf{x}_{1:t}, \mathbf{y}_{1:t}\}$ with observations distributed as $y_t \sim \mathcal{N}(f(\mathbf{x}_t), \sigma^2)$. Let $[\mathbf{\Phi}]_{i:} = \boldsymbol{\phi}(\mathbf{x}_i)$ denote the features evaluated at the observed inputs. Given this data the posterior the weights will be normally distributed $\boldsymbol{\theta}|\mathcal{D}_t \sim \mathcal{N}(\mathbf{A}^{-1}\mathbf{\Phi}^\mathsf{T}\mathbf{y}_{1:t}, \sigma^2\mathbf{A}^{-1})$ where $\mathbf{A} = \mathbf{\Phi}^\mathsf{T}\mathbf{\Phi} + \sigma^2\mathbf{I}$. Since the feature weights are normally distributed it must hold that that $\boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\boldsymbol{\theta}$ will be as well, and as a result we can obtain the posterior of $f(\mathbf{x})|\mathcal{D}_t$ as a Guassian with mean and variance given by

$$\mu_t(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\mathbf{A}^{-1}\mathbf{\Phi}^\mathsf{T}\mathbf{y}_{1:t}, \tag{3}$$

$$\sigma_t^2(\mathbf{x}) = \sigma^2\boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\mathbf{A}^{-1}\boldsymbol{\phi}(\mathbf{x}). \tag{4}$$

Equivalently we can rewrite these quantities in terms which only make use of inner products between the features, i.e.

$$\mu_t(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\mathbf{\Phi}^\mathsf{T}(\mathbf{\Phi}\mathbf{\Phi}^\mathsf{T} + \sigma^2\mathbf{I})^{-1}\mathbf{y}_{1:t} \tag{5}$$

$$\sigma_t^2(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\mathbf{\Phi}^\mathsf{T}(\mathbf{\Phi}\mathbf{\Phi}^\mathsf{T} + \sigma^2\mathbf{I})^{-1}\mathbf{\Phi}\boldsymbol{\phi}(\mathbf{x}). \tag{6}$$

As a result we can see that this linear model approximates the full GP introduced earlier.

References

S. Bochner. *Lectures on Fourier integrals.* Princeton University Press, 1959.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Neural Information Processing Systems*, pages 1177–1184, 2007.

C. E. Rasmussen and C. K. Williams. *Gaussian processes for machine learning.* The MIT Press, 2006.