On the Equivalence between Quadrature Rules and Random Features

Francis Bach

Arthur Gretton's notes

October 23, 2015

What the paper is about

The paper has two parts:

1. Approximating functions by random Fourier features is similar to Herding (and more generally, quadrature).
2. A non-uniform sampling distribution can improve performance of both methods.

What the paper is about

The paper has two parts:

1. Approximating functions by random Fourier features is similar to Herding (and more generally, quadrature).
2. A non-uniform sampling distribution can improve performance of both methods.

Outline of this talk:

- Approximate RKHS functions by random Fourier features (review)
- Introduce what's meant by quadrature (approximating integrals)
- Show the quadrature problem is not, in fact, equivalent
- Probably not covered: the non-uniform sampling distribution

Function approximation by random Fourier features

Reminder: Fourier representation of RKHS. Kernel

$$k(x, y) = k(x - y),$$

Fourier series representation of k, for $\mu_\ell \geq 0$,

$$\begin{aligned} k(x - y) &= \sum_{\ell=0}^{\infty} 2\hat{k}_\ell \left[\cos(\ell x)\cos(\ell y) + \sin(\ell x)\sin(\ell y)\right] \\ &= \sum_{\ell=0}^{\infty} \mu_\ell \varphi(\ell, x)\varphi(\ell, y) \end{aligned}$$

E.g. "Gaussian-like" kernel:

$$k(x - y) = \frac{1}{2\pi}\vartheta\left(\frac{(x - y)}{2\pi}, \frac{\imath\sigma^2}{2\pi}\right), \qquad \mu_\ell = \frac{1}{\pi}\exp\left(-2\sigma^2 \lfloor l/2 \rfloor^2\right).$$

ϑ is the Jacobi theta function, close to Gaussian when σ^2 sufficiently narrower than $[-\pi, \pi]$.

Function approximation by random Fourier features

Functions are in RKHS iff they can be written wrt a function $g \in L_2(\mu)$,

$$f(x) = \sum_{\ell=0}^{\infty} \underbrace{[\sqrt{\mu_\ell} g_\ell]}_{f_\ell} \underbrace{[\sqrt{\mu_\ell} \varphi(\ell, x)]}_{\phi_\ell(x)} \qquad \sum_{\ell=0}^{\infty} \mu_\ell g_\ell^2 < \infty$$

Function approximation by random Fourier features

Functions are in RKHS iff they can be written wrt a function $g \in L_2(\mu)$,

$$f(x) = \sum_{\ell=0}^{\infty} \underbrace{[\sqrt{\mu_\ell} g_\ell]}_{f_\ell} \underbrace{[\sqrt{\mu_\ell} \varphi(\ell, x)]}_{\phi_\ell(x)} \qquad \sum_{\ell=0}^{\infty} \mu_\ell g_\ell^2 < \infty$$

Approximate the function f, for $v_i \in \mathbb{N}$ and $\alpha_i \in \mathbb{R}$,

$$\hat{f} = \sum_{i=1}^{n} \alpha_i \varphi(v_i, \cdot) \in \widehat{\mathcal{F}}.$$

Error is (for some reference measure ρ)

$$\left\| \hat{f} - f \right\|_{L_2(\rho)} = \left\| \sum_{i=1}^{n} \alpha_i \varphi(v_i, x) - \sum_{\ell=0}^{\infty} \mu_\ell g_\ell \varphi(\ell, x) \right\|_{L_2(\rho)}.$$

Simplest case: $v_\ell \overset{\text{i.i.d.}}{\sim} \mu$ and $\alpha_\ell = n^{-1} g(v_\ell)$. Then $\mathbb{E} \left\| \hat{f} - f \right\|_{L_2(\rho)}^2 \leq n^{-1} C$.

Can we do better?

Quadrature definition

What is quadrature? Approximate the integral $\int_{\mathcal{X}} h(x)g(x)d\rho(x)$ via

$$\sum_{i=1}^{n} \alpha_i h(x_i) - \int_{\mathcal{X}} h(x)g(x)d\rho(x)$$

for $\alpha \in \mathbb{R}^n$ and $x_1, \ldots, x_n \in \mathcal{X}$, and $h \in \mathcal{F}$ an RKHS function, ρ a prob. measure, for some

$$g \in L_2(\mathcal{X}).$$

Quadrature definition

What is quadrature? Approximate the integral $\int_{\mathcal{X}} h(x)g(x)d\rho(x)$ via

$$\sum_{i=1}^{n} \alpha_i h(x_i) - \int_{\mathcal{X}} h(x)g(x)d\rho(x)$$

for $\alpha \in \mathbb{R}^n$ and $x_1, \ldots, x_n \in \mathcal{X}$, and $h \in \mathcal{F}$ an RKHS function, ρ a prob. measure, for some

$$g \in L_2(\mathcal{X}).$$

KEY POINT: for RKHS, approximating the integral can be done by approximating a function. For $\|h\|_{\mathcal{F}} \leq 1$,

$$\left|\sum_{i=1}^{n} \alpha_i h(x_i) - \int_{\mathcal{X}} h(x)g(x)d\rho(x)\right| = \left|\left\langle h, \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) - \int_{\mathcal{X}} k(x, \cdot)g(x)d\rho(x)\right\rangle_{\mathcal{F}}\right|$$

$$\leq \left\|\sum_{i=1}^{n} \alpha_i k(x_i, \cdot) - \int_{\mathcal{X}} k(x, \cdot)g(x)d\rho(x)\right\|_{\mathcal{F}}.$$

When $g(x) = 1$ this is Herding, since $\mu_\rho = \int_{\mathcal{X}} k(x, \cdot)d\rho(x)$.

Quadrature definition

To implement quadrature, approximate the function

$$\int_{\mathcal{X}} k(x,\cdot)g(x)d\rho(x) \in \mathcal{F}$$

by the function

$$\sum_{i=1}^{n} \alpha_i k(x_i,\cdot) \in \mathcal{F}$$

Ensure the error small in RKHS norm.

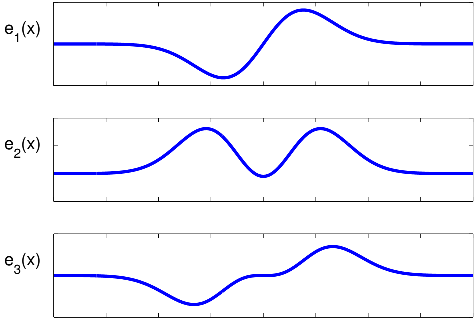
1. Can we make this look like the random Fourier feature loss? (in a manner of speaking, after a math detour)
2. Simplest case: $x_i \overset{\text{i.i.d.}}{\sim} \rho$ and $\alpha_i = n^{-1}g(x_i)$. Then $\mathbb{E}\left\|\sum_{i=1}^{n}\alpha_i k(x_i,\cdot) - \int_{\mathcal{X}} k(x,\cdot)g(x)d\rho(x)\right\|_{\mathcal{F}} \leq n^{-1}C$. Can we do better?

RKHS in terms of eigenfunctions of integral operator

Gaussian kernel, $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$,

$$\begin{aligned} \lambda_k &\propto b^k \qquad b < 1 \\ e_k(x) &\propto \exp(-(c-a)x^2) H_k(x\sqrt{2c}), \end{aligned}$$

a, b, c are functions of σ, and H_k is kth order Hermite polynomial.



$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x')$$

RKHS in terms of eigenfunctions of integral operator

Define an integral operator with the kernel k and probability distribution ρ:

$$\begin{aligned} T_k f \ &: L_2(\rho) \to L_2(\rho) \\ & f \mapsto \int k(x,t) f(t) d\rho(t) \end{aligned}$$

The eigenfunctions of the kernel with respect to some measure ρ are

$$\lambda_i e_i(x) = \int k(x,t) e_i(t) d\rho(t) = T_k e_i$$

We can prove $\sum_i \lambda_i < \infty$ and $\lambda_i \geq 0$ (normalizable).

$$k(x,x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x'), \qquad \int_{\mathcal{X}} e_i(x) e_j(x) d\rho(x) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

Under certain conditions (e.g Mercer's) this sum is guaranteed to converge absolutely and uniformly (whatever the x and x').

RKHS in terms of eigenfunctions of integral operator

Define the RKHS using the eigenfunctions:

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x') = \left\langle \phi(x), \phi(x') \right\rangle_{\mathcal{F}}$$

Infinite dimensional feature map: $\phi(x) = \left[\ \dots \ \ \sqrt{\lambda_i} e_i(x) \ \ \dots \ \right] \in \ell_2.$

RKHS function: $\forall \{f_i\}_{i=1}^{\infty} \in \ell_2.$

$$f(x) = \sum_{i=1}^{\infty} f_i \phi_i(x) = \sum_{i=1}^{\infty} f_i \sqrt{\lambda_i} e_i(x)$$

RKHS in terms of eigenfunctions of integral operator

Define the RKHS using the eigenfunctions:

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$$

Infinite dimensional feature map: $\phi(x) = \left[\ \dots \ \ \sqrt{\lambda_i} e_i(x) \ \ \dots \ \right] \in \ell_2.$

RKHS function: $\forall \{f_i\}_{i=1}^{\infty} \in \ell_2.$

$$f(x) = \sum_{i=1}^{\infty} f_i \phi_i(x) = \sum_{i=1}^{\infty} f_i \sqrt{\lambda_i} e_i(x)$$

For this to work, the dot product in $\mathcal{F}$ must be

$$\langle f, g \rangle_{\mathcal{F}} = \sum_{i=1}^{\infty} f_i g_i = \left\langle T_k^{-1/2} f, T_k^{-1/2} g, \right\rangle_{L_2(\rho)}$$

In other words $\|f\|_{\mathcal{F}}^2 = \sum_i f_i^2 = \left\| T_k^{-1/2} f \right\|_{L_2(\rho)}^2$.

RKHS in terms of eigenfunctions of integral operator

Start with a function $g \in L_2(\rho)$, expanded in terms of the basis $e_i(x)$,

$$g = \sum_{i=1}^{\infty} \langle g, e_i \rangle_{L_2(\rho)} e_i.$$

Then obtain a function $f \in \mathcal{F}$ via

$$f(x) = T_k^{1/2} g = \sum_{i=1}^{\infty} \underbrace{\langle g, e_i \rangle_{L_2(\rho)}}_{f_i} \sqrt{\lambda_i} e_i(x).$$

since $\sum_{i=1}^{\infty} \langle g, e_i \rangle_{L_2(\rho)}^2 = \|g\|_{L_2(\rho)}^2 < \infty$.

RKHS in terms of eigenfunctions of integral operator

Start with a function $g \in L_2(\rho)$, expanded in terms of the basis $e_i(x)$,

$$g = \sum_{i=1}^{\infty} \langle g, e_i \rangle_{L_2(\rho)} e_i.$$

Then obtain a function $f \in \mathcal{F}$ via

$$f(x) = T_k^{1/2} g = \sum_{i=1}^{\infty} \underbrace{\langle g, e_i \rangle_{L_2(\rho)}}_{f_i} \sqrt{\lambda_i} e_i(x).$$

since $\sum_{i=1}^{\infty} \langle g, e_i \rangle_{L_2(\rho)}^2 = \|g\|_{L_2(\rho)}^2 < \infty$.
Also possible for the kernel:

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x') = T_k^{1/2} \left(\sum_{i=1}^{\infty} \sqrt{\lambda_i} e_i(x) e_i(x') \right) = T_k^{1/2} \psi(x, x').$$

The final result

We can write the function approximation as a loss in $L_2(\rho)$:

$$\begin{aligned}
&\left\|\sum_{i=1}^{n}\alpha_i k(x_i,\cdot)-\int_{\mathcal{X}}k(x,\cdot)g(x)d\rho(x)\right\|_{\mathcal{F}}\\
&=\left\|\sum_{i=1}^{n}\alpha_i T_k^{1/2}\psi(x_i,\cdot)-\int_{\mathcal{X}}T_k^{1/2}\psi(x,\cdot)g(x)d\rho(x)\right\|_{\mathcal{F}}\\
&=\left\|\sum_{i=1}^{n}\alpha_i\psi(x_i,\cdot)-\int_{\mathcal{X}}\psi(x,\cdot)g(x)d\rho(x)\right\|_{L_2(\rho)}.
\end{aligned}$$

The final result

We can write the function approximation as a loss in $L_2(\rho)$:

$$\begin{aligned}
&\left\| \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) - \int_{\mathcal{X}} k(x, \cdot) g(x) d\rho(x) \right\|_{\mathcal{F}} \\
&= \left\| \sum_{i=1}^{n} \alpha_i T_k^{1/2} \psi(x_i, \cdot) - \int_{\mathcal{X}} T_k^{1/2} \psi(x, \cdot) g(x) d\rho(x) \right\|_{\mathcal{F}} \\
&= \left\| \sum_{i=1}^{n} \alpha_i \psi(x_i, \cdot) - \int_{\mathcal{X}} \psi(x, \cdot) g(x) d\rho(x) \right\|_{L_2(\rho)} .
\end{aligned}$$

Reminder: random Fourier problem was

$$\left\| \hat{f} - f \right\|_{L_2(\rho)} = \left\| \sum_{i=1}^{n} \alpha_i \varphi(v_i, \cdot) - \sum_{\ell=0}^{\infty} \mu_\ell g_\ell \varphi(\ell, x) \right\|_{L_2(\rho)} .$$

Main difference: spatial vs frequency decomposition.